


Explaining and Predicting Helpfulness and Funniness of Online Reviews on the Steam Platform

Zhi Wang, Xi'an Jiaotong-Liverpool University, Suzhou, China

Victor Chang, Teesside University, Middlesbrough, UK

 <https://orcid.org/0000-0002-8012-5852>

Gergely Horvath, Duke Kunshan University, Suzhou, China

ABSTRACT

The online review is a crucial display of many online shopping platforms and an essential source of product information for consumers. Low-quality reviews often cause inconvenience to the platform and review readers. This article aims to help Steam, one of the largest digital distribution platforms, predict the review helpfulness and funniness. Via Python, 480,000 game reviews related data for 20 games were captured for analysis. This article analyzed the impact of three categories of influencing factors on the usefulness and funniness of game reviews, which are characteristics of review, reviewer, and game. Additionally, by using the random forest-based classifier, the usefulness of reviews could be accurately predicted, while for funniness, gradient boosting decision tree was the better choice. This article applied research on the usefulness of reviews to game products and proposed research on the funniness of reviews.

KEYWORDS

Sentiment Analysis, uGame Analysis, User Feedback Review

1. INTRODUCTION

Online user reviews have gradually become an important display content of many online shopping platforms with the full application of Web 2.0 and the rise of social networking sites and new media. It refers to consumer's judgment (evaluation, response, recommendation, or complaint) on a product or service after purchase on the Internet. These reviews contain a large amount of useful information, and because the data comes from different individuals, it can avoid the one-sidedness of information. An online review is one of the most important influencing factors for customers' decision-making when buying unfamiliar products. Meanwhile, from the perception of business firms, it contains various user feedback on products or services, which can provide manufacturers with Research & Development and improvement strategies. User reviews have increasingly become the focus of theoretical and social practice. Recently, the amount of consumer reviews has been increasing rapidly, creating significant big data challenges for consumers and businesses (Singh et al., 2017).

It is essential for online businesses to continually improve and even innovate their products from user experience analysis. Online review, an important component of user-generated content, is an excellent source for research and commercial use and it shows the feasibility of getting quality

DOI: 10.4018/JGIM.20211101.aa16

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

improvement tips from it (Yang et al., 2019). For example, through gamers' comments, game developers could discover the advantages and disadvantages of games and bugs they did not notice before. After that, they can release a new version to fix and update the game according to information obtained from reviews. It also can give them directions in developing new games because of the understanding of users' needs. Nowadays, the development of emotional analysis technology and text mining technology has greatly improved the ability to perceive customer needs and provide timely feedback.

For consumers, online user reviews give them an opportunity to know more about products before making decisions. These reviews were generated by customers who have purchased and used the products. This is a good way to reduce information asymmetry in market transactions. However, due to too many reviews, the amount of information is much higher than what the customers can consume, withstand, or need. It is impossible for a potential buyer to read through all the reviews of the product. In most cases, customers refer to the first few comments displayed on the product review page, so only comments that consumers have read are likely to help them make decisions. If helpful reviews are more visible to potential buyers, it will be easier for buyers to get the useful information they need (Singh et al., 2017). Therefore, it encourages online retailers to evaluate the helpfulness of reviews. Many websites now set a question next to consumer reviews, like "Is this comment helpful to you?" with "thumbs up" and "thumbs down" buttons.

The helpfulness of reviews is usually just assessed by the number of helpfulness voting. Most websites rank reviews in the order according to the usefulness of reviews voted by consumers and put the review that gets the most votes in the first. It helps consumers to reduce the cost of information search and increase the efficiency of purchasing decisions. By improving the user-friendliness of the website, more users will be attracted. However, there exist some problems in identifying the value of reviews in this way. Firstly, the latest user reviews may be placed after the previously less useful ones since too many comments are posted at the same time, so that cannot be voted in time. Secondly, for some products that are not so popular, almost no one proposes for the usefulness, so these reviews cannot be ranked by usefulness. These two reasons will both cause the failure of the voting mechanism. Consequently, it is necessary to research factors affecting the helpfulness of reviews and build prediction models for online businesses to forecast helpfulness of any new reviews and sort them effectively.

Many previous studies did research on online user reviews of various product categories, such as movies, phones, restaurant services, hotels, clothes and others. As determinant factors of review helpfulness might differ for different products, this article will focus on the game reviews on a digital game social platform, Steam platform, with little research in this area.

The 'Steam platform' is one of the largest comprehensive digital distribution platforms for PC gaming in the world, developed by Valve Corporation. Up to now, Steam's operation has been very successful. Numerous game issuers have released and updated their games on this platform. It has around 555 million active users in total and over 12,000 games in the database (Galyonkin, 2019). Through the Steam Store and the Steam Community, Steam provides digital rights management, multiplayer games and social networking services (Lin et al., 2019). Users can purchase, download, discuss, upload and share games and software on the platform.

Players are allowed to make comments on games they played on the game's store page. Unlike the popular five-star rating system adopted by most platforms, Steam uses a simpler binary rating system, asking users to provide an overall impression of the game: "Recommended" or "Not Recommended". In addition to recommendations and reviews, more information is shown on the page: number of owned games, number of reviews posted, number of playing hours of the game reviewed, number of people who found the review helpful or funny. Users can edit reviews after posting them, but only one comment can be made on a game. Users viewing these comments can give feedback by answering the question: "Was this review helpful" and there are three choices: Yes, No and Funny. Steam provides a brief summary of the game's reviews for each game on its store page, showing the

percentage of positive user reviews (Recommended) and the number of total reviews. The trends of reviews posted are displayed by graph and a sudden increase in new comments may result from the release of new updates.

We chose to work on game reviews as the game industry is now a mature and rapidly developing industry. PC games grew by 4.0% to \$35.7 billion year-on-year in 2019. Despite the small size and slow growth of the PC game segment market, its position as the innovation basis of game market is still obvious (Tom Wijman, 2019). It is beneficial for game developers to understand the concerns of gamers who are difficult to satisfy. In addition, the development of PC games brings the trend of game platformization and a large number of third-party game platforms have emerged. From a global perspective, the steam platform has the longest history and is the largest industry leader. In 2018, Steam had an average of 47 million active users per day and 90 million active users per month (Fenlon, 2019). Thus, studying game reviews on this platform will lead to representative conclusions.

This article will focus on two main problems: **1) How can the influencing factors affect perceived helpfulness of review? 2) Predict the helpfulness of online customer reviews on the Steam platform and find out which machine learning method can build a more accurate helpfulness prediction model.**

For solving the problems, 480,000 game reviews are collected by Python, including data of review text, recommendation, helpful votes, funny votes and other variables. After data preprocessing, the review characteristics are extracted from the collected data. Next, regression analysis is applied to examine the influencing factors. The results show positive and negative relationships between helpfulness and factors, which would be explained in the following article. Finally, by changing the dependent variable into a binomial variable, two data mining methods (Random Forest & Gradient Boosting Decision Tree) are implemented to predict review helpfulness and funniness. After evaluating, it can be found that Random Forest performed better for helpfulness prediction while GBDT got a higher accuracy for funniness.

The outline of this paper is as follows. First, Section 4 discusses the related work and is followed by the introduction of the methodology in Section 5. The steps of data collection, preprocessing are shown in Section 6, along with the data overview. The models used and results are explained in Section 7. Finally, Section 8 & 9 separately presents the evaluation of results and conclusions.

2. LITERATURE REVIEW

The concept of the helpfulness of online reviews was first proposed by Chatterjee(Chatterjee, 2001), referring to the degree of influence of the use of review information. Nowadays, the review usefulness has been widely used to measure the value of comments to improve consumers' ability to evaluate product quality. It is regarded as a subjectively perceived value of whether online reviews are helpful in consumers' decision-making process (Mudambi and Schuff, 2010). It is a standard for measuring the quality of online reviews. In particular, appropriate information about product quality can be contained in user reviews to create more value for potential consumers (Zhou and Yang, 2019). Therefore, the usefulness of online reviews has attracted the attention of many scholars and has become a research hotspot in the field of online reviews. As the usefulness of comments is considered to be the subjective perception of the reader, measuring the usefulness of comments is still in the exploratory stage. Some scholars used questionnaires to obtain it, while most scholars measured usefulness based on the percentage of online review votes on the total votes on the website. Many studies collected product review information from the Amazon website and used a helpful vote rate to conduct research. Singh et al. approximated the suitable ratio by using each review's percent convenience, which is useful votes divided by total votes attracted by all reviews written by this reviewer (Singh et al., 2017).

Studies on online review helpfulness mainly focused on the influencing factors and helpfulness prediction. The following introduces mostly related research in these two aspects.

2.1 Factors Influencing the Helpfulness of Online Reviews

There are many aspects affecting online review helpfulness that were involved in previous researches, such as the characteristics of review (linguistic characteristics, content), reviewer (reputation) and product (product type).

2.1.1 *Characteristics of Review*

In terms of review attributes, it refers to the characteristics of the review itself, including review star rating, review content, review time, etc. A review star score is a reviewer's evaluation of a product in terms of the number of stars, ranging from one to five stars. Most scholars found negative reviews more useful than positive ones. Liu and Park (2015) came to the opposite conclusion: positive reviews are considered more valuable by consumers than negative and neutral reviews.

Comment content is an important comment attribute, usually analyzed from the perspective of text characteristics (review length & readability) and sentiment features (polarity, subjectivity). Review length is measured in the review's word count. It is generally believed that the more words a review has, the more information it contains, thus it is more helpful. Mudambi & Schuff (2010) verified that the word count of a review had a significant positive impact on the usefulness of the review regardless of search goods or experience goods. However, in Racherla and Friske's (2012) research, word count might be an unimportant factor.

Ghose and Ipeirotis (2011) studied the impact on usefulness from the perspective of the readability. They believed that the choice of review words, misspellings would affect readability and thus affect usefulness. It was proved that there was a positive correlation between readability and usefulness. Agnihotri and Bhattacharya (2016) concluded that there was a nonlinear relationship between readability and helpfulness with diminishing marginal effect. Krishnamoorthy (2015) considered various other linguistic features, like adjectives, state verbs, action verbs, and subjectivity. This article used a hybrid set of four kinds of features (review metadata, subjectivity, readability, and linguistic category) to build a model. Malik and Hussain (2018) studied the impact of review content and reviewer characteristics on review usefulness and found that the number of space, aux verb, drives words in review text are essential predictors for review helpfulness.

Posting time reflects how long the review has been posted, which is also an influencing factor. Cao (2011) believed that the posting time is inversely related to the usefulness of the comment. Additionally, Wang et al. (2015) found that its different effects on search and experience products. Early posted reviews of experience products would receive more helpful votes, while search products would receive less. However, some researchers hold the opposite view that the longer the comment is posted, the more useful the comment would be (Lee, 2013).

2.1.2 *Characteristics of the Reviewer*

Recently, more descriptive information about the reviewer has been prominently displayed on the online retailer's website. The characteristics of the reviewer generally include personal information disclosure, the professionalism and the reputation of the reviewer, which reflects the credibility of the reviewer.

Some reviewers will disclose their personal information such as names, hobbies, photos on the website. Many researchers agreed that such behavior has a positive effect on review helpfulness (Ghose and Ipeirotis, 2011). Racherla and Friske's (2012) research did not support that. They also researched the impact of reviewer expertise on review usefulness, where the expertise of a reviewer was judged by the number of comments posted in the past and the content of the comments. They concluded that reviews written by highly professional reviewers are more useful for message recipients. However, Liu and Park (2015) found that the variables of expertise were not significantly related to the usefulness of the assessment. One of the possible explanations would be the different motivations of online review readers.

2.1.3 Characteristics of Product

Product characteristics also include an important influencing factor that many studies emphasize. Products are mainly classified as two types: search and experience products, according to consumer's ability to get product information before buying (Nelson, 1970). Most consumers' perception of search products, such as mobile phones and digital cameras, can be obtained through a web search for the main objective quality parameters. Unlike search products, it is difficult to make a buying decision for experience goods by checking parameters. Product quality can only be accessed from personal experience and the reviews would be more subjective. Examples of experience goods include games, books and movies. Many studies added product types to the research on the impact of online review helpfulness or tried to find the different impact of factors in different types. Mudambi and Schuff (2010) conducted empirical research on these two types of products, proving that the type of products significantly influenced review usefulness. Chua and Banerjee (2016) analyzed the moderating role of product type in usefulness. They found the presence of expertise claims in experience product reviews had a significant relationship with review helpfulness while not for experience products.

2.2 Predictions of Online Review Helpfulness

The focus of predictions on the usefulness of online reviews is primarily the accuracy. Whether reviews are useful can be identified by establishing a regression model or a classification model depending on whether the dependent variable is numeric or nominal. When the favorable ratio is set as the dependent variable, linear regression is the most used model due to its fast speed and explanation capability. Therefore, many previous studies used linear regression to predict the review helpfulness score. However, more scholars chose other models in their research, such as Tobit regression (Mudambi & Schuff, 2010; Korfiatis, 2012; Agnihotri and Bhattacharya, 2016), Zero Inflated regression. Mudambi & Schuff (2010) also analyzed why Tobit regression is better than OLS regression in his study. There are also researchers who build predictive models based on neural networks. Lee and Choeh (2014) applied a back-propagation multilayer perceptron neural network (BPN) model for prediction and found it has better accuracy than linear regression.

Some studies convert the favorable raw ratio to nominal data like 'helpful' or 'unhelpful' based on whether the raw percentage exceeds a baseline threshold. In this case, the problem is transformed into a classification problem. Ghose and Ipeirotis (2011) used Random Forest classifiers to solve the problem and found that the classifiers performed better in search of products than experienced ones. In Krishnamoorthy's research (2015), Naive Bayes, Support Vector Machine and Random Forest were tested as learning methods by using 10-fold cross-validation. After evaluating predictive models through f-measure and accuracy, Random Forest outperformed the other methods. Hu and Chen (2016) applied three methods (SVR, linear regression, and Model tree (M5P)) on the TripAdvisor dataset to predict usefulness and showed M5P significantly performed best. Additionally, Park (2018) examined SVR, Linear Regression, Random Forest and M5P on Amazon.com datasets, determining that the best method was support vector regression. Thus, as for which method can better predict the review helpfulness, different studies have different conclusions.

In summary, it can be known that in addition to using traditional statistical methods, data mining methods such as machine learning in the computer field are popular to be used to make predictions.

3. METHODOLOGY

The methodologies used in this article are introduced in this section. The first is a data processing and data extraction, which would be explained in Section 6. Secondly, exploratory data analysis would be applied to view the overall situation of the data. Thirdly, the Poisson regression and Zero Inflated Poisson regression model will be built to find how the dependent variables affect a dependent variable. Finally, the helpfulness and funniness are predicted by using two methods: Random Forest

and Gradient Boosting Decision Tree. The performance of these two models should be compared using confusion matrix, ROC curve and AUC.

3.1 Zero Inflated Poisson Regression Model (ZIP)

In this dissertation, the zero-inflated Poisson regression model is applied to analyze how the influencing factors affect review helpfulness. ZIP Regression model combines these two regression models, taking advantage of them for processing this type of data. It regards the data set as a mixture of a data set that is all 0 and a not-always-zero data set that satisfies the Poisson distribution. Which dataset each observation in is determined by the result of a Bernoulli trial.

When it comes to regression analysis, the method of Ordinary Least Squares (OLS) is most commonly used. Since the dependent variable and the most independent variables do not meet the normal distribution, OLS regression may cause a huge deviation. It is known that the dependent variable in this article is count data. For the processing of such data, some classic discrete models are used, such as the Poisson model, Negative binomial model, generalized Poisson model and others. Additionally, the helpful votes count contains a high proportion of zero value because most of the reviews were never viewed and voted. The zero-inflated model is suitable for data with many zero values and over-dispersion values. The estimation results have strong validity, so reliable hypothesis tests and parameter estimation can be obtained.

However, given the variance of the dependent variable is much greater than the mean, showing that over-dispersion existed. Poisson regression requires the relationship between prediction variables and response variables to follow a Poisson distribution with equal average and dispersion (Fitriani et al., 2019). After data transformation, the dependent variable would be changed into data with equal average and dispersion, so applying this regression model is reasonable.

3.2 Machine Learning Techniques

Besides the regression model, this article chooses some widely used machine learning methods to predict the helpfulness of online game reviews. There are Random Forest and Gradient Boosting Decision Tree. As each technique has its own pros and cons, finding the most suitable model for the data set is important. After applying each method and comparing the results using some evaluation methods, it will be determined. Here is the necessary information about these ML techniques.

3.2.1 *Random Forest*

Random Forest is an algorithm that integrates multiple decision trees through the idea of ensemble learning. It randomly selects k features in a decision tree with m features to form n decision trees and then selects the prediction result. For regression problems, the decision tree in a random forest will predict Y 's value (the output value). The final predicted value is calculated by averaging the predicted values of all decision trees in Random Forest.

Random Forest can deal with classification and numerical characteristics simultaneously, so it can be used to solve classification and regression problems. Besides, Random Forest has strong adaptability to the data and does not need the normalization of the data set. It can also reduce the risk of overfitting by obtaining the average. However, due to its complexity, it is more time-consuming to train than other similar algorithms.

3.2.2 *Gradient Boosting Decision Tree*

The Boosting method is a method that combines a series of weak prediction models (usually decision trees) to generate prediction models, thereby improving the accuracy of the weak classification algorithm. The boosting method uses a stage-wise way to build the model. The weak learner built at each step of the iteration is to make up for the shortcomings of the existing model. A boosting tree is an iteration of multiple regression trees to make collective decisions. The core of GBDT is

that each tree learns the residuals of the sum and conclusion of all the trees before and fits a current residual regression tree.

GBDT can get higher accuracy with relatively less tuning time. It has a wide range of uses that can flexibly process various types of data, including continuous and discrete values. However, there are dependencies between weak learners so that it is difficult to train data in parallel.

3.3 Model Evaluation Indicator

After the model construction is completed, the effect of the model needs to be evaluated. Based on the evaluation results, the performance of different models should be compared. The simplest and most used indicator for evaluating a model is accuracy, but it often does not reflect the performance of a model. When comparing the performance of different models, using different indicators leads to different evaluation results. Due to the different training requirements of the model, the indicators for evaluating the performance of the model will also differ. In this article, the purpose of modeling is to find useful comments with few misjudgments. Common indicators are confusion matrix, ROC curve and AUC (Area Under the ROC Curve).

3.3.1 Confusion Matrix

When dealing with classification problems, a confusion matrix is a kind of index that comprehensively reflects the performance of the model and many indexes can be derived from it. Table 1 shows the matrix.

Table 1. Confusion Matrix

| Confusion Matrix | | True Class | |
|------------------|----------|----------------------|----------------------|
| | | Positive | Negative |
| Prediction Class | Positive | TP True Positive | FP False Positive |
| | Negative | FN False Negative | TN True Negative |

In the face of a large number of data, it is difficult to measure the quality of the model just by calculating the number. Therefore, the confusion matrix extends the following four indicators in the basic statistical results:

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- Accuracy = $(TP + TN) / (TP + TN + FP + FN)$
- F1-score = $(2 * Recall * Precision) / (Recall + Precision)$

The F1-Score indicator combines the results of Precision and Recall. The value of F1-Score ranges from 0 to 1, where 1 represents the best model output and 0 represents the worst.

3.3.2 ROC Curve and AUC

The AUC indicator is used in the binary classification problem. The model evaluation phase is often used as the most important evaluation indicator to measure the stability of the model. The ROC curve can be obtained by plotting the True Positive Rate (TPR) as the vertical axis and the False Positive Rate (FPR) as the horizontal axis, while AUC is the area under the ROC curve. The value of AUC

ranges from 0.5 to 1; 0.5 corresponds to the “random guessing model” on the diagonal. Generally, the better the model, the larger its AUC.

The ROC curve and AUC are used because the ROC curve has a very good characteristic: when the distribution of the positive and negative samples in the test set changes, the ROC curve can remain unchanged. However, under the background of category imbalance, the large number of negative cases makes FPR growth not obvious, resulting in an overly optimistic effect estimation of the ROC curve.

4. DATA/DEVELOPMENT

4.1 Data Collection

The game review data used in this study was collected from the Steam website using Python. As Steam has a ‘STEAM & GAME STATS’ page showing top games by current player count, I chose the top 20 most-played games on 17th Sep, which could represent recent trends. Analyzing their reviews can lead to some representative conclusions.

As the Steam platform divided users into different regions according to their nationality regions, and published games according to user zones, some of the games on the list were not sold in China, so their comments cannot be searched and collected. We skipped these kinds of games. Moreover, for secure processing, we only collected reviews in the English version. One more thing that should be mentioned is that although these games are all popular recently, the numbers of reviews are not on the same quantity level because of different release dates. For example, for CS, a game released in 2012, it has more than 1 million reviews in English. It is difficult to collect and analyze all of them, so about 5% of them were collected. For games with 60 thousand reviews, half of them were collected. The principle of selecting comments was to select comments with helpfulness evaluation as much as possible and further filtering will be done in the data preprocessing.

Review data collected had ten attributes: **recommendation** (Whether the reviewer recommended the game or not); **game’s name**; **review_posted** (When the review was posted); **helpful** (How many players thought the review was helpful); **funny** (How many players thought the review was funny); **hour_played** (How many hours a reviewer played the game till now), **is_early_access_review**; **received_for_free**; **review** (Content of the review); **number_of_games** (How many games the reviewer had in his/her account). Furthermore, some game information could be collected manually in the game’s store page, such as the **game’s release date**, **price**, **categories** the game belongs to, **positive rate** (percentage of positive reviews) and **English review** (number of English reviews) and **Total review** (number of reviews in total).

We first found the top 20 games’ website addresses manually and wrote a program by the Python package ‘Beautiful Soup’ to collect data separately into 20 text files. Beautiful Soup is a simple way to extract data from HTML or XML files with very few code lines. After getting 20 text documents, we imported them into Excel separately to have an intuitive inspection and understanding of data and used Excel’s own data processing tools to remove duplicates. Finally, we had one document with a raw data size of 172 MB and 480 thousand records. We also created another Excel worksheet containing game information collected manually.

4.2 Data Pre-Processing

As the original data was sometimes incomplete and inconsistent, data mining could not be carried out directly. The results were unsatisfactory, and data preprocessing technology should be used to improve the quality of data analysis. Before starting, it is necessary to integrate two documents together using the function ‘merge’ to merge them on the ‘Name’ column.

The first step is to handling missing or misplaced values. Table 2 shows the number of missing values in each variable. It can be reasonably concluded that the missing values in ‘Received_for_free’ and ‘Is_early_access_review’ mean ‘It was not received for free’ and ‘It was not early access review’

Table 2. Number of missing values in each variable

| Column name | Number of Missing values |
|------------------------|--------------------------|
| Received_for_free | 469055 |
| Is_early_access_review | 441713 |
| Number_of_games | 381 |
| Review | 94 |
| Hour_played | 26 |

respectively, so we assigned them a value of 'No'. For 'Number_of_games' or 'Hour_played', there were few cases that no value was collected. We assumed the reviewer got only one game in his game library or spent so little time playing the game that did not be recorded by Steam, so I set the value as 1 and 0. We also removed records with no words in the content of the review. Finally, the number of records reduced to 479917 and there was no missing value.

The second step is to extract numerical data from text information, including categorical encoding, as most of the machine learning algorithms require numerical input. For 'Recommendation', 'Is_early_access_review' and 'Received_for_free', I encoded them by creating a dictionary with the mapping given as 1: Yes and 0: No. The data of the number of people found this review helpful and funny was combined in one sentence in raw data, so we used some methods to extract numbers into two new variables: 'helpful' and 'funny'.

The third step is to handle outliers. By roughly viewing the output of function 'describe' which showing count, mean, maximum, and other basic information for each numerical column. We found that the 'funny' column had some values more than 4 billion, which was unreasonable and might be caused by an error when gripping data. As the error only appeared on a tiny part of records, we chose to drop all of them. Meanwhile, there were negative numbers in the 'date' column (shows the number of days between the comment released and the game released), which was counter-intuitive. One of the reasons was that the review was an early access review written by players who played the game before it was officially released. Another exception might be that Steam was not the first distribution platform for the game. Some reviewers bought games from other platforms and posted comments on Steam. Whether this part of data should be taken into consideration would be decided later based on analysis results.

As helpful or funny, votes would be not only affected by the review's helpfulness or interestingness but also by the popularity of games. Reviews of games with fewer players would apparently get fewer votes. This will cause the difference between the values of different games' reviews, making the data analysis results inaccurate. It is necessary to scale the data in a certain proportion so that it falls in a specific area convenient for comprehensive analysis. The specific operation will be mentioned in the modeling section later (Figure 1).

In order to ensure consistency of data, through review's post date, together with the game's release date, the number of days between the time of comment release and game release can be calculated. Additionally, the English review rate can be computed by dividing the number of English reviews by total reviews.

4.3 Data Overview

Through preliminary data processing, some basic impressions on these top 20 games could be achieved. 479,763 valid data were collected in total, in which 321,909 reviews had 'helpful' or 'funny' feedbacks.

In terms of the game, the 20 games are in four genres: Action (Adventure), Strategy, RPG and Simulation. Figure 2 shows the number of games in each genre. Counter-Strike: Global Offensive, PLAYERUNKNOWN'S BATTLEGROUNDS and Team Fortress 2 have the most comments. From

Figure 1. Descriptive data

| | Recommendation | helpful | funny |
|-------|----------------|------------------------------|------------------------------|
| count | 305243.000000 | 305243.000000 | 305243.000000 |
| mean | 0.707875 | 10.091488 | 3.940487 |
| std | 0.454740 | 144.806547 | 87.947005 |
| min | 0.000000 | 1.000000 | 0.000000 |
| 25% | 0.000000 | 1.000000 | 0.000000 |
| 50% | 1.000000 | 1.000000 | 0.000000 |
| 75% | 1.000000 | 3.000000 | 0.000000 |
| max | 1.000000 | 28185.000000 | 15569.000000 |

the proportion of English comments in all comments, it can be concluded that Risk of Rain 2, Sid Meier's Civilization® V and Team Fortress 2 are more attractive to English language users because they have more English reviews. Most games collected have more than 80% positive feedbacks.

In terms of reviews, 77.7% of reviewers said they would recommend the games they played in the entire database. 75% of reviews got less than three helpful votes and got 0 funny votes, showing many comments have not received any feedback on their helpfulness and funniness. Meanwhile, some reviews had relatively high helpful votes, which indicates that most consumers were not keen on voting for the usefulness of online reviews unless the review was particularly valuable to them. According to the popularity of the game, the number of votes received varies greatly. Most reviewers played the game for less than 600 hours and had less than 200 games in their game library (Figure 3).

For sentiment analysis, the three graphs below show the polarity distribution, subjectivity calculated by Textblob and compound (polarity) by vaderSentiment. Comparing the polarity computed by different methods, they reached completely different conclusions. From Textblob, it shows that most reviews' polarity concentrated in the middle area and the mean is 0.11, representing neutral or slightly positive sentiment in most reviews. However, the compound score has an average of 0.33 and more than 65 percent of reviews have a compound score over 0.05, denoting positive sentiments. Though

Figure 2. Genres of games

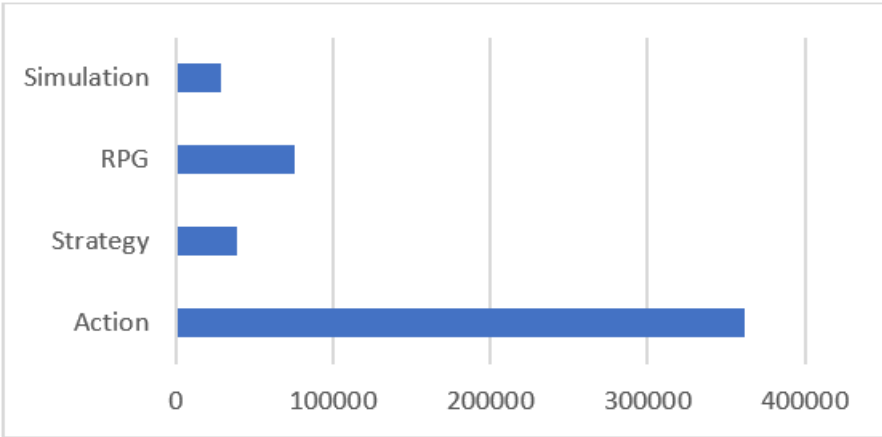
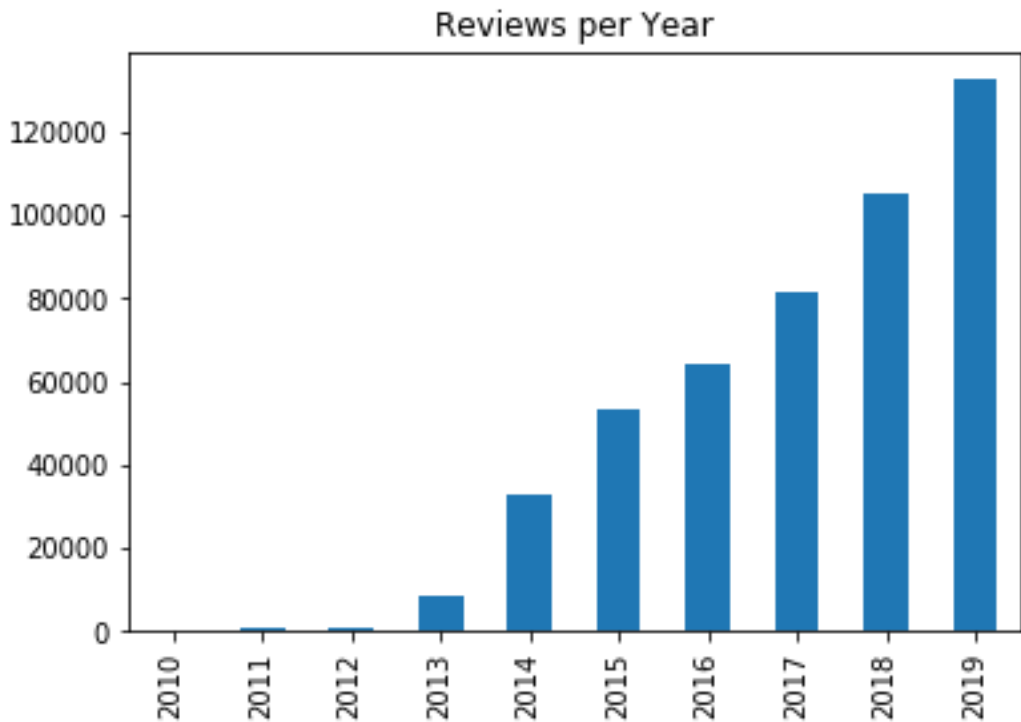


Figure 3. Review per Year



it is said that Vader Sentiment Analysis works better for with texts from social media, compound and polarity are both used in the model.

In terms of readability, from the calculation of Flesch Reading Ease, it can be found that the mean is 65.5, meaning the reviews are plain English. It is worth noting that some reviews with particularly low scores with a minimum of -676340 could affect the mean. This happened because some reviewers repeatedly wrote some same words without spaces (like 'WARMONGERWARMONGERWARMONGERWARMONGER.....') or repeated the last letter of the word many times (like 'uhhhhhhhhhhhhhhhhhhh') to express their excitement. Excluding those values, less than -300, the mean of readability would increase to 78. It can be concluded that gamers tend to comment on the game in a language that is fairly easy to understand (Figure 4).

5. MODEL AND RESULTS

As mentioned in the methodology part, Poisson and Zero Inflated Poisson regression model is applied to analyze the influence of different dependent variables on review helpfulness and funniness rated by community members in this dissertation. A prediction modeling will then be taken after changing the helpful variable into a logical variable with values 0 and 1 (0 for helpless and 1 for helpful). This becomes a binary classification problem and some classification methods were used. In this section, the explanatory econometric analysis and prediction results of some machine learning methods will be presented.

Figure 4a. Sentiment Analysis: Polarity, Subjectivity and Compound

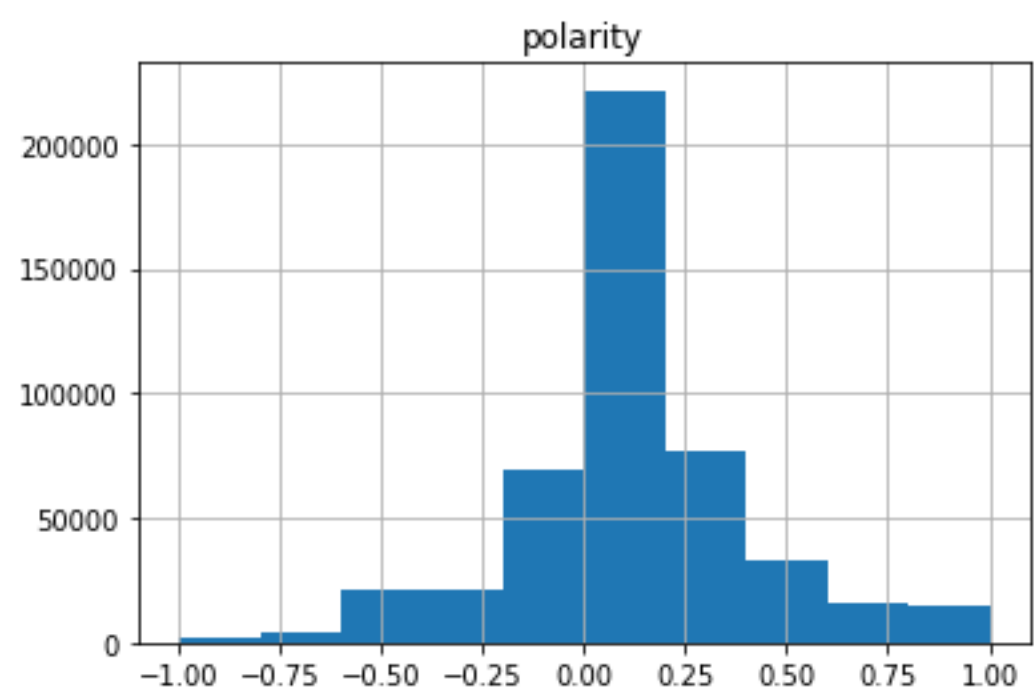


Figure 4b. Sentiment Analysis: Polarity, Subjectivity and Compound

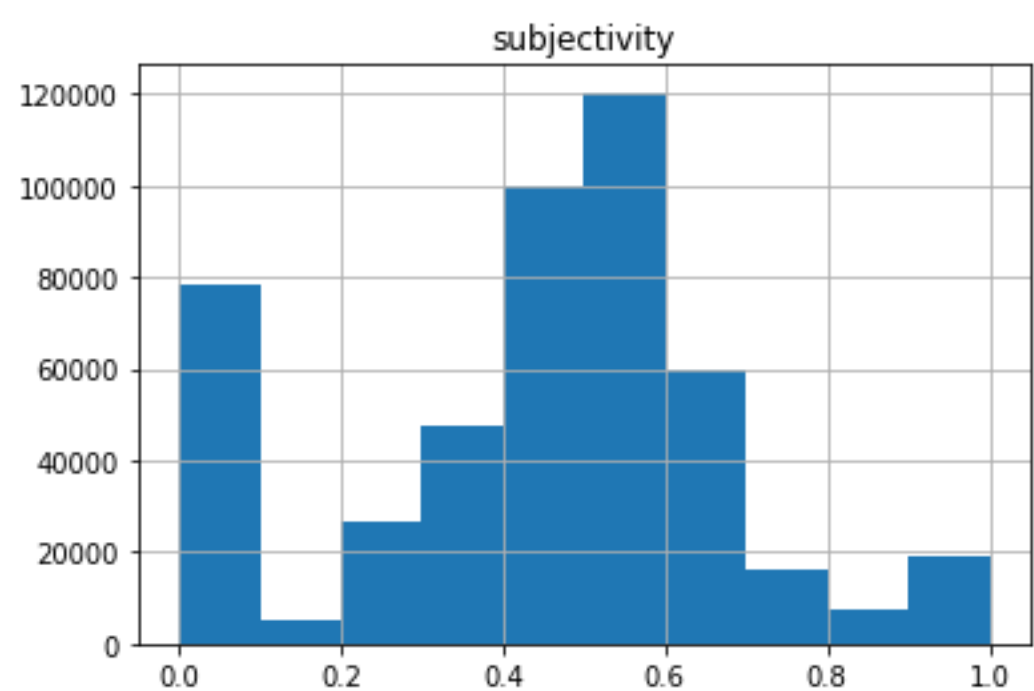
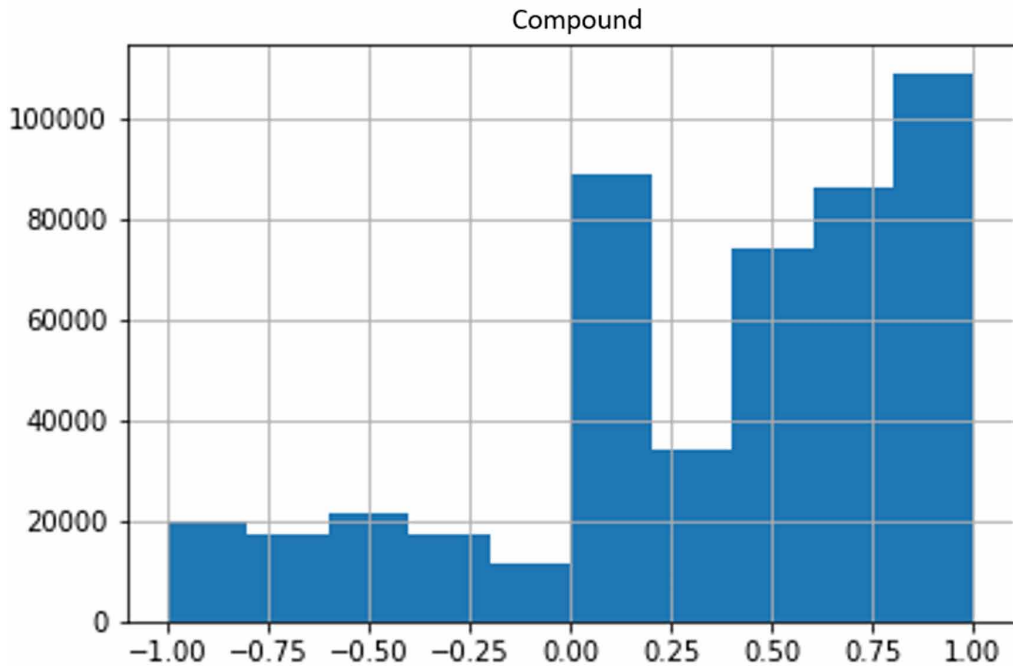


Figure 4c. Sentiment Analysis: Polarity, Subjectivity and Compound



5.1 Variables

The dependent variable in this article is to review helpfulness. In prior researches based on the dataset collected from Amazon.com, the helpfulness ratio was calculated by the ratio of the number of helpful votes to the total number of votes (Korfiatis et al., 2012). The Steam website did not provide total votes number, so I used the count number of helpfulness votes.

The explanatory variables for review usefulness prediction can be divided into four categories: a review (exclude review text), review text, reviewer and game. Below are dependent variables initially determined to be used in the model. These would be adjusted after processing.

1. Properties of review
 - **Recommendation:** Reviewer's overall view of the game: recommended or not
 - **Elapsed_days:** Number of days between the review posted and collected
 - **is_early_access_review:** Whether the review was an early-access review (The review was for the trial version of the game)
2. Review text

Review content is an important review attribute that most researchers focused on, and it can generally be analyzed from the perspective of text features such as comment length and comment readability. Variables chosen in this article can be categorized into three groups: linguistic, psychological, and readability.

Linguistic variables:

- **word_count**: The number of words in the review text
- **Noun**: The number of nouns in the review text
- **Adjective**: The number of adjectives in the review text
- **Verb**: The number of verbs in the review text
- **Stop_word**: The number of stop words in the review text

Psychological variables:

- **Polarity (& compound)**: Tendentious emotion in the review text: positive or negative, ranging from -1.0 to 1.0. (-1.0 for negative and 1.0 for positive)
- **Subjectivity**: Whether the review text expresses personal feelings, views, or beliefs. It ranges from 0.0 to 1.0. (0.0 for objective, 1.0 for subjective)

For measuring polarity and subjectivity of the review text, Textblob is a tool that can be applied. It is an open-source text processing library written in Python. Textblob can be used to perform many natural language processing tasks, such as speech tagging, nominal component extraction, sentiment analysis, and text translation. When calculating polarity and subjectivity, Textblob looked for words or phrases that can calculate their sentiment and average them over longer texts (Schumacher, 2015).

Additionally, VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis can be used to calculate emotional polarity (Hutto and Gilbert, 2014). It uses manual annotation (10 people) to determine the sensitive polarity and intensity of 7000+ commonly used emotional words (including adjectives, nouns, adverbs, etc.). Unlike other sentiment dictionaries, VADER 's dictionaries also consider widely used characters (such as:)) and abbreviations to deal with sentiment discrimination of non-standard sentences in the network environment. When analyzing comments or texts from social media, the sentiment of sentences changes according to emoticons. Vader takes this into account with slang, capital letters, etc., so it is a better choice for social media review analysis and sentiments. VADER provides a single scoring criterion called the vaderSentiment compound. The 'compound' can be regarded as 'polarity' in Textblob in the same range [-1, 1].

Readability variables:

Readability can be considered a measure of the ease of reading or understanding text, indicating how effectively the writer's meaning is conveyed to the reader. Comments that are easy to read by a large number of users will also get more votes (Ghose and Ipeirotis, 2011). There are numerous readability formulas for measuring, such as Flesch Reading Ease formula, Flesch-Kincaid Grade Level, Fog Scale (Gunning FOG Formula), SMOG Index and others. This article selected one method:

- **Flesch Reading Ease:**

It is considered to be one of the oldest and most accurate readability formulas. The specific calculation formula is:

$$RE = 206.835 - (1.015 * ASL) - (84.6 * ASW)$$

Where,

RE = Readability Ease

ASL = Average number of words per sentence (number of words divided by the number of sentences)

ASW = Average number of syllables per word

The output, RE, falls between 0 and 100 in most cases. The higher the score, the higher the readability. It can be calculated through 'textstat' package. Theoretically, the value has a maximum of 121.22 and there is no limit on the minimum value. (Aggarwal, n.d.)

3. Reviewer

The characteristics of commentators are also considered to be a major factor. In terms of game reviews, reviewers' playtime and the number of games they own can reflect their reliability. This dissertation selected three indicators that can be collected:

- **hour_played**: Number of hours the reviewer played the game till now
- **release_day**: Number of days between the review posted and game released
- **number_of_games**: Number of games the reviewer owned in his/her account

4. Game

- **positive_rate**: Percentage of positive reviews
- **total_review**: Number of all comments on the game
- **English_review_rate**: Proportion of English comments
- **game_release_day**: Number of days since the game was released

5.2 Factors Determining Review Helpfulness

The first research question is concerned with how the review or reviewer's characteristics affect the perceived helpfulness of review. Poisson model and Zero Inflated Poisson model are tested and similar results are found. Data used for modeling excluded comments with zero votes because these reviews are probably not read by the reader. The number of votes cannot represent their helpfulness or funniness.

5.2.1 Model Specification

We first performed a correlation test analysis to diagnose multicollinearity between independent variables. It was found that there was strong multicollinearity among word count, stop words, verb, noun and adjective, which is intuitive. The moderate relationships between 'Recommendation' and 'Positive rate', 'Recommendation' and 'compound' could also be discovered. In the subsequent analysis, we tried to remove those highly co-linear variables, but finally, we found that there was not much impact on the model construction. This will be presented in the 'empirical results' part.

Before performing a regression analysis, the data should be transformed to meet the normative requirements for easy mining. Otherwise, the model's convergence speed and model accuracy will be affected. As the distribution of the dependent variable and most independent variables such as 'helpful', 'funny', 'Number_of_games', etc. was approximately heavy-tailed. These variables have more distribution on smaller values and big values that are important and cannot be discarded. This leads to uneven distribution and huge variance. Thus, we used log transformation on these variables to stabilize the variance. Log transformation is useful when applied to skewed distributions because it tends to stretch the range of independent variable values that fall within a lower range and compress the range of variable costs that fall within a lower range. For other variables like 'Game_release_day' and 'Elapsed_days', their distribution is not dispersed, but the range is relatively large. In this case, I scaled the data appropriately.

5.2.2 Empirical Results

Next, the Poisson Regression and Zero Inflated Poisson Regression were performed to examine the explanatory variables that affect usefulness. Figures 5-7 separately present results through different models and different variables.

From the comparison of Figures 2 and 3, it can be found that removing these variables (stop words, verb, noun & adjective) has no significant effect on the regression analysis of the entire model. Moreover, Poisson Regression and Zero Inflated Poisson Regression got basically consistent results.

Despite 'Received_for_free', 'stopwords' and 'verb', other variables were considered significant for the regression model. The recommendation has a negative impact on review helpfulness, indicating that reviews that do not recommend this game instead are more informative. Similarly, reviews of

Figure 5. Poisson Regression Results

| | coef | std err | z | P> z | [0.025 | 0.975] |
|------------------------|---------|---------|----------|-------|--------|--------|
| Recommendation | -0.7033 | 0.006 | -113.234 | 0.000 | -0.715 | -0.691 |
| funny | 0.4528 | 0.001 | 302.657 | 0.000 | 0.450 | 0.456 |
| Is_early_access_review | 0.0586 | 0.011 | 5.403 | 0.000 | 0.037 | 0.080 |
| Received_for_free | 0.0197 | 0.017 | 1.129 | 0.259 | -0.014 | 0.054 |
| Hour_played | 0.0288 | 0.001 | 19.384 | 0.000 | 0.026 | 0.032 |
| Number_of_games | 0.1049 | 0.002 | 52.969 | 0.000 | 0.101 | 0.109 |
| Date | -0.0165 | 0.001 | -17.248 | 0.000 | -0.018 | -0.015 |
| Positive rate | -0.6700 | 0.022 | -31.021 | 0.000 | -0.712 | -0.628 |
| English/Total | -0.2038 | 0.030 | -6.733 | 0.000 | -0.263 | -0.144 |
| Elapsed_days | -0.2370 | 0.020 | -12.064 | 0.000 | -0.275 | -0.198 |
| Game_release_day | -0.8059 | 0.021 | -37.729 | 0.000 | -0.848 | -0.764 |
| polarity | -0.0404 | 0.011 | -3.709 | 0.000 | -0.062 | -0.019 |
| subjectivity | -0.0904 | 0.012 | -7.602 | 0.000 | -0.114 | -0.067 |
| word_count | 0.0865 | 0.006 | 15.344 | 0.000 | 0.075 | 0.098 |
| compound | 0.0323 | 0.005 | 6.988 | 0.000 | 0.023 | 0.041 |
| stopwords | -0.0024 | 0.005 | -0.441 | 0.659 | -0.013 | 0.008 |
| verb | 0.0070 | 0.005 | 1.407 | 0.160 | -0.003 | 0.017 |
| noun | -0.0135 | 0.006 | -2.403 | 0.016 | -0.025 | -0.002 |
| adjective | 0.0273 | 0.005 | 5.319 | 0.000 | 0.017 | 0.037 |
| Readability | -0.0091 | 0.001 | -7.204 | 0.000 | -0.012 | -0.007 |
| Review/Date | -0.3116 | 0.011 | -28.231 | 0.000 | -0.333 | -0.290 |

Figure 6. Poisson Regression Results (variables removed)

| | coef | std err | z | P> z | [0.025 | 0.975] |
|------------------------|---------|---------|----------|-------|--------|--------|
| Recommendation | -0.7025 | 0.006 | -113.275 | 0.000 | -0.715 | -0.690 |
| funny | 0.4525 | 0.001 | 303.553 | 0.000 | 0.450 | 0.455 |
| Is_early_access_review | 0.0559 | 0.011 | 5.155 | 0.000 | 0.035 | 0.077 |
| Received_for_free | 0.0211 | 0.017 | 1.212 | 0.226 | -0.013 | 0.055 |
| Hour_played | 0.0288 | 0.001 | 19.423 | 0.000 | 0.026 | 0.032 |
| Number_of_games | 0.1044 | 0.002 | 52.883 | 0.000 | 0.101 | 0.108 |
| Date | -0.0171 | 0.001 | -18.136 | 0.000 | -0.019 | -0.015 |
| Positive rate | -0.6763 | 0.021 | -31.575 | 0.000 | -0.718 | -0.634 |
| English/Total | -0.2133 | 0.030 | -7.083 | 0.000 | -0.272 | -0.154 |
| Elapsed_days | -0.2469 | 0.019 | -12.667 | 0.000 | -0.285 | -0.209 |
| Game_release_day | -0.7999 | 0.021 | -37.568 | 0.000 | -0.842 | -0.758 |
| polarity | -0.0342 | 0.011 | -3.160 | 0.002 | -0.055 | -0.013 |
| subjectivity | -0.0925 | 0.012 | -7.854 | 0.000 | -0.116 | -0.069 |
| word_count | 0.0952 | 0.002 | 53.513 | 0.000 | 0.092 | 0.099 |
| compound | 0.0332 | 0.005 | 7.191 | 0.000 | 0.024 | 0.042 |
| Readability | -0.0104 | 0.001 | -8.377 | 0.000 | -0.013 | -0.008 |
| Review/Date | -0.3178 | 0.011 | -29.472 | 0.000 | -0.339 | -0.297 |

games with a higher positive rate seem to contain less helpful information. For characteristics of reviewers, both 'Hour_played' and 'Number_of_games' positively affect helpfulness while 'Date' (number of dates from a game released and review posted) has the opposite effect. In other words, reviews posted by experienced gamers who own more games and spend more time playing this game are more useful for reviewers. Further, comments that were posted right after the game launched contain more new information and can attract more votes. It is noteworthy that game release day and elapsed days have a negative relationship with the helpfulness of reviews, meaning that new games attract more useful reviews. This may indicate a trend that the quality of reviews has increased recently. In terms of review content, it is known that a review comprising many words, objective and little more complicated, is perceived as useful by readers.

5.3 Helpfulness Prediction

The final research question is to train a model to predict the helpfulness of game reviews and compare the accuracy of each method. We decided to change the count value of helpful votes into a binary variable and used various machine learning approaches for classification. This section will first show the methodology for changing the problem into a binary classification problem and present the results of used algorithms.

5.3.1 Binary Classification Problem

Unlike most other researches that set helpfulness ratio (helpful votes/ total votes of each review) as a dependent variable, the helpfulness in my data set is helpful votes by review readers. Therefore, the variable was highly dispersed and limited by the popularity of the game, compared with that helpfulness ratio ranging from 0 to 1. It is difficult to predict the exact number. I tried to make predictions with regression analysis and some data mining methods like Support Vector Regression and the outcome was not accurate enough or time-consuming.

For simplicity, changing it to a binary classification problem would be a reasonable choice. A threshold α can be set to divide the 'useful' and 'useless' reviews. In detail, reviews with more than α helpful votes were marked as useful and others as not useful. Due to the different number of helpful votes attracted by various games, α should also be selected differently. Assumed that each game has the same percentage of helpful reviews, the percentage rate could be found and applied to find α in each game. It is crucial to find out the suitable value for the useful review percentage for precise separation.

Inspired by the method used in Ghose and Ipeirotis's (2011) article, this dissertation took similar steps. I asked two human coders for the content analysis of 500 reviews by asking them, 'Is this review useful or not?'. The main aim is to find a useful proportion of reviews. The 500 reviews were selected randomly regardless of games. The coders had no access to the number of helpful votes in case they were affected by the judgment of other readers. They could see the name of the game the review referred to and whether the reviewer recommended it, which was helpful for their judgment. Finally, the analysis showed that the conclusions of the two were basically the same. Approximately 23.9% of the reviews were considered useful from their work.

There was some misclassification. For example, some comments with many votes were considered useless by humans, or some with few votes were considered useful. Ghose and Ipeirotis (2011) performed a ROC analysis to find the optimal threshold that minimizes the error rates. After adjustment, they changed their threshold from 0.739 from 0.6. Back to this dissertation, due to the relatively high proportion of low votes, the slight change in the percentage had little impact on the number of votes set as the threshold. Thus, we chose 24% as the boundary value. In other words, we assumed that 24 percent of reviews were useful for each game and classified them as 'Helpful'. Others were marked as 'Unhelpful'.

5.3.2 Model Building

Before applying different supervised learning techniques, the total data set should be divided into two parts: the training set and testing set by using 'train_test_split' module from a Python package. 80% of the dataset was randomly kept in the training set to train and estimate the model and the remaining 20% was for testing the performance of the optimal model selected. Since the total number of samples was adequate, the k-fold cross-validation method did not need to be adopted.

Although Support Vector Machine classification was considered to work well in most binary classification problems, its main disadvantage is the difficulty of being implemented on large training samples. Since SVM uses the quadratic programming to solve the support-vector, solving the quadratic programming will involve calculating a matrix of order m (m is the number of samples). When the number of m is large, the storage and calculation of the matrix will consume a lot of machine memory and operation time. We standardized the dataset and experimented with the SVM method, finding that the running time was too long to draw conclusions. Thus, we used Random Forest and Gradient Boosting Decision Tree.

Note that the data set is imbalanced, meaning the number of samples of one class is much smaller than other classes in the data set. In this case, most (76%) reviews were regarded in the same type (helpless). It often leads to the classifier's output tending to the majority category in the data set: outputting the majority category will bring higher classification accuracy, but it will perform poorly in the minority category we are concerned (helpful reviews). There are some ways to deal with this problem: resampling (oversampling & undersampling), an ensemble of sampler (Badr, 2019) and threshold shift. Undersampling is a method to maintain a reasonable proportion of data in each category by filtering out some data from the biggest data category. Due to a large amount of data, the method was applied in the testing set and increased the model performance in predicting the 'helpful' category.

5.3.3 Empirical Results

Through a series of basic tuning parameters, the two methods were examined and the evaluation results were presented below. When using random forest, we also tested the results without processing the input data with resampling. Comparing the classification reports and accuracy scores, random forest without resampling achieved a higher accuracy score. The recall rate of data predicted to be 0 (helpless) is 98% and that of data predicted to be helpful is 35%. The huge difference indicated that classifiers did bias towards category with large sample size. After balancing the data and remodeling, the recall rate for useful reviews increased from 35% to 76% and the precision dropped. In total, the F1-score improved slightly. The comparison of the two results revealed that after undersampling, the trained model could find more useful comments, but at the same time, the precision will be appropriately reduced. Whether this change will help in real-life situations remains to be discussed.

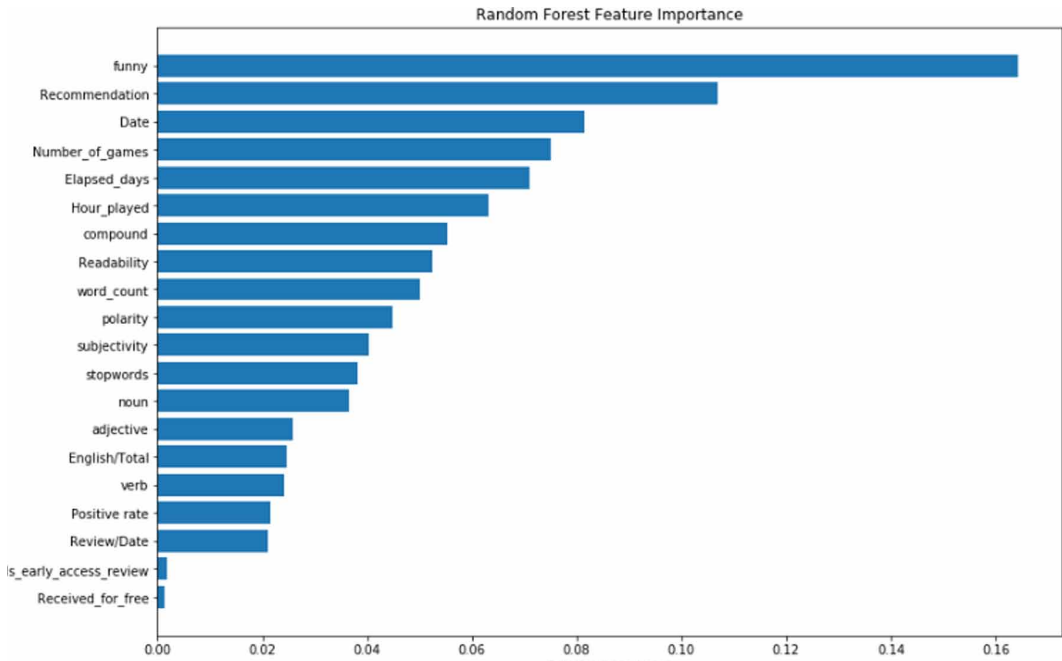
Comparing the Random Forest and Gradient Boosting Decision Tree's results, there was a tiny difference. Figure _ presents ROC curves and AUC values, showing RF's curve is slightly above GBDT's. In general, the index indicates RF and GBDT have similar model accuracy while RF performs slightly better in this case.

In addition, after building and testing the classification model, feature importance for each input variable could be calculated. From the Random Forest Feature Importance (Figure 5), we found that the funniness influenced the perceived helpfulness most. In a sense, more interesting comments will also contain more useful information and will attract more people for useful votes. The influencing variable 'Recommendation' was also important. Since games are experience goods, the reviewer's recommendation that expressed his general view occupies an essential position in the classification.

The high importance for 'Date' shows that comments posted time can help voters make judgments easier. For 'Number of games' and 'Hour played', the relatively high variable importance indicates that reviewers' experience would be an important determinant for readers to vote positively or negatively. In addition, variables such as 'Positive rate', 'Review/Date', 'Is_early_access_review' and

‘Received_for_free’ did not make much impact on customers when voting for usefulness. Feature importance computed by GBDT also achieved similar results (Figure 7).

Figure 7. Empirical Results and Analysis by Random Forest



5.4 Factors Determining Review Funniness and Funniness Prediction

The same steps used in usefulness analysis are also applied to funniness. The result of Zero Inflated Poisson Regression of review funniness is shown in Tables 3 and 4. Comparing it with helpfulness, they have both similarities and differences.

Unlike the result of helpfulness regression, ‘Recommendation’ has an adverse effect on funniness, which means recommended comments are sometimes more interesting. This result makes sense because people usually praise rather than criticize in interesting language. Another big difference is that ‘Hour_played’ and ‘Number_of_games’ influence negatively on funniness. Game reviewer’s experience makes the game review less interesting. Additionally, with a positive correlation between word count and funniness, it is surprising that stop words, verbs, and adjectives have a negative relationship with funniness. Other variables are basically similar to the results in terms of helpfulness.

Before running the prediction model, changing funny votes into a binary variable needed to be considered, unlike helpfulness, determining if a comment is interesting is subjective, so the results of human coding are quite different. I lowered the threshold to 60% for determining interesting.

After taking the same steps and adjusting parameters, classification reports and ROC curves were obtained. The classification results showed these two methods have different advantages. Random Forest has better accuracy, but GBDT has a higher recall rate in funniness. From the ROC curve, GBDT performed better, which is different from what concluded from helpfulness prediction.

Table 3. Prediction for No-Votes Reviews for Example 1

| Recommendation | Funny | Hour played | # of games | Date | Elapsed days | Polarity | Compound | Subjectivity | Word count | Stop words | Readability |
|----------------|-------|-------------|------------|------|--------------|----------|----------|--------------|------------|------------|-------------|
| 0 | 0 | 602.4 | 21 | 111 | 293 | -0.01 | 0.9662 | 0.56 | 81 | 36 | 69.45 |

5.5 Prediction for No-Votes Reviews

In this section, we applied the trained random forest model to comments that did not get any helpful votes. The main reason these comments were not voted on was that they were covered by others, so the model can automatically give them a helpful judgment and help the website reorder reviews. As in this case, only helpful reviews are essential; several examples of comments that are judged ‘useful’ are shown below.

Example 1: ‘(Monster Hunter) Single-player is the preferred way to play because online matchmaking with friends completely sucks. The game seems to think it’s fair to kick me offline for being in a quest alone or standing idle for the 30s. Friends do not even show up when trying to join Friends Sessions, and even when by some miracle we can join them, the game kicks either my friend or myself after a mission or two (if we’re lucky and the game didn’t kick us immediately)’.

Example 2: (Warframe) ‘Only have a few hours into this game. It is pretty fun so far! I can’t wait to see what they do with this game in the future! Good looter shooter type game, they let you feel the power fantasy, but do try to keep it from going to far down the rabbit hole. Devs are about as open as they can be, with bi-weekly twitch streams on progress, and are responsive to the community, a more than refreshing breeze in the recent years. Yes, it does have microtransactions, but there only a few things (prime access accessories) that can’t be got from in-game grind. There is an open trade system within the game where you can sell certain items to other players to get other loot or the premium currency (Platinum), which is how you can even get all but the above mentioned cosmetic items. Pretty well-balanced system IMO’.

Although previous model results indicated that the accuracy of finding useful reviews was not high, these examples seem reasonable.

6. EVALUATION

The interpretation of the regression coefficients and the results of the prediction model provides some further insights into review helpfulness and funniness, mostly in the aspect of determining factors and suitable classification methods. The findings reached are generally consistent with those in previous studies. In detail, from the analysis above, my main findings were shown in the following.

First, we found that ‘Not recommended’ reviews are more useful for review readers, probably because people often point out the game’s disadvantages when they show their dissatisfaction, which contains more information. It is similar to the findings in most literature. However, Korfiatis et al. (2012) held the opposite view that high helpfulness is affected by the positive rating value. They thought consumers tend to read reviews that support this product first. We also concluded that the expertise of a reviewer (can be represented by played time and number of games owned in this article)

Table 4. Prediction for No-Votes Reviews for Example 2

| Recommendation | Funny | Hour played | # of games | Date | Elapsed days | Polarity | Compound | Subjectivity | Word count | Stop words | Readability |
|----------------|-------|-------------|------------|------|--------------|----------|----------|--------------|------------|------------|-------------|
| 1 | 0 | 2477.8 | 108 | 2072 | 295 | -0.007 | 0.96 | 0.54 | 149 | 77 | 66.61 |

has a positive effect on review usefulness. This also confirms the conclusions of other researches. In contrast, there is an interesting finding that the reviewer's expertise is negatively related to review funniness. One guess is that more experienced gamers tend to post more professional reviews, which loses fun. Additionally, the significant impact of review length suggests longer reviews often increase the usefulness of reviews, consistent with the findings of Lee and Choeh (2017). Further, the number of reviews 'elapsed' days has a negative relationship with review helpfulness, which is the same as Cao's (2011) results. It is also worth noting that subjective reviews seem to be helpless and less funny. This conclusion is a bit counter-intuitive and may be due to a biased and subjective calculation.

For prediction methods, it is found that Random Forest gives the best results for helpfulness prediction, sharing a similar conclusion with Krishnamoorthy (2015), and Gradient Boosting Decision Tree performed better for funniness prediction. Other studies achieved different conclusions for the best algorithms, so suitable methods should be different for different data sets and variables.

This study derives the exploration of the usefulness of reviews into the field of game reviews. By analyzing how these factors affect the perceived helpfulness of review, a better understanding of review and review helpfulness in the game area can be obtained. As consumer reviews have become an important source of product information, game websites also need to understand how their clients perceive reviews. In addition, based on my findings, the classification model can help the website to identify reviews that are expected to be helpful or funny for readers and rank them first when displaying. To some extent, it can avoid the issue that new posted useful review is not visible because of no helpful votes. Meanwhile, the article mentioned a new review perceived attribute, funny. It can provide readers with some thoughts on how to analyze it deeper.

7. CONCLUSION AND RECOMMENDATIONS

In this paper, two main research questions were solved. First, we constructed a model to evaluate the different determinants of review helpfulness and funniness. The results showed that word count, the number of games, hours played had a positive relationship with review helpfulness. In contrast, the review's elapsed day, review release day, subjectivity, and readability negatively impacted it. For review funniness, the variables, recommendation, number of games, and hours played were drawn to the opposite effect as helpfulness. Finally, among the two widely used classification methods, Random Forest performed better for helpfulness prediction while the ensemble learning technique GBDT achieved a higher accuracy. This study would help the Steam website identify helpful reviews even if reviews had little or no manual voting.

There are several limitations to this study. First, only English reviews were selected and analyzed in this article, a part of all reviews. In the Steam website, some games have a large number of Chinese and Japanese players, so the conclusions have no guiding effect on ranking the comments generated by these players. Second, we did not verify the subjectivity, polarity and compound computed by Textblob and vaderSentiment. Whether these tools are suitable for analyzing game reviews should be researched in the future. Finally, this article provided an overall discussion of the interestingness of studies as a supplement to review helpfulness but did not explain it in detail. Our future research will extend the analysis to multiple languages for reviews and will explore how to measure and influence review helpfulness in detail.

ACKNOWLEDGMENT

This work is partly supported by VC Research (VCR 0000102).

REFERENCES

- Aggarwal, S.B., & Chaitanya. (n.d.). *Textstat: Calculate statistical features from text*. Academic Press.
- Agnihotri, A., & Bhattacharya, S. (2016). Online Review Helpfulness: Role of Qualitative Factors: Online Review Helpfulness. *Psychology and Marketing*, 33(11), 1006–1017. doi:10.1002/mar.20934
- Badr, W. (2019). *Having an Imbalanced Dataset? Here Is How You Can Fix It*. Data Sci. <https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb>
- Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach. *Decision Support Systems*, 50(2), 511–521. doi:10.1016/j.dss.2010.11.009
- Chatterjee, P. (2001). Online Reviews: Do Consumers Use Them? *Assoc. Consum. Res.*, 6.
- Chua, A. Y. K., & Banerjee, S. (2016). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior*, 54, 547–554. doi:10.1016/j.chb.2015.08.057
- Fenlon, W. (2019). Steam now has 90 million monthly users. *PC Gamer*. <https://www.pcgamer.com/steam-now-has-90-million-monthly-users/>
- Fitriani, R., Chrisdiana, L. N., & Efendi, A. (2019). Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data. *IOP Conf. Ser. Mater. Sci. Eng.*, 546. doi:10.1088/1757-899X/546/5/052025
- Galyonkin, G. (2019). SteamSpy - All the data about Steam games. *SteamSpy - Data Steam Games*. <https://steamspy.com/>
- Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 1498–1512. doi:10.1109/TKDE.2010.188
- Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management*, 36(6), 929–944. doi:10.1016/j.ijinfomgt.2016.06.003
- Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth Int. Conf. Weblogs Soc. Media*, 10.
- Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3), 205–217. doi:10.1016/j.elerap.2011.10.003
- Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7), 3751–3759. doi:10.1016/j.eswa.2014.12.044
- Lee, J. (2013). What Makes People Read an Online Review? The Relative Effects of Posting Time and Helpfulness on Review Readership. *Cyberpsychology, Behavior, and Social Networking*, 16(7), 529–535. doi:10.1089/cyber.2012.0417 PMID:23742150
- Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6), 3041–3046. doi:10.1016/j.eswa.2013.10.034
- Lee, S., & Choeh, J. Y. (2017). Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees. *Management Decision*, 55(4), 681–700. doi:10.1108/MD-06-2016-0398
- Lin, D., Bezemer, C.-P., Zou, Y., & Hassan, A. E. (2019). An empirical study of game reviews on the Steam platform. *Empirical Software Engineering*, 24(1), 170–207. doi:10.1007/s10664-018-9627-4
- Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, 47, 140–151. doi:10.1016/j.tourman.2014.09.020
- Malik, M., & Hussain, A. (2018). An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing & Management*, 54(1), 88–104. doi:10.1016/j.ipm.2017.09.004

- Mudambi, S., & Schuff, . (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *Management Information Systems Quarterly*, 34(1), 185. doi:10.2307/20721420
- Nelson, P. (1970). Information and Consumer Behavior. *Journal of Political Economy*, 78(2), 311–329. doi:10.1086/259630
- Park, Y.-J. (2018). Predicting the Helpfulness of Online Customer Reviews across Different Product Types. *Sustainability*, 10(6), 1735. doi:10.3390/su10061735
- Racherla, P., & Friske, W. (2012). Perceived ‘usefulness’ of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, 11(6), 548–559. doi:10.1016/j.elerap.2012.06.003
- Schumacher, A. (2015). *TextBlob Sentiment: Calculating Polarity and Subjectivity*. https://planspace.org/20150607-textblob_sentiment/
- Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Kumar Roy, P. (2017). Predicting the “helpfulness” of online consumer reviews. *Journal of Business Research*, 70, 346–355. doi:10.1016/j.jbusres.2016.08.008
- Wang, T., Wang, K., Chen, H. (2015). The Impact of Temporal Distance on the Increase of the Perceived Usefulness of Online Reviews: From the Perspective of the Attribution Theory. *J. Bus. Econ.*, 2, 46–56. doi:10.3969/j.issn.1000-2154.2015.02.006
- Wijman, T. (2019). The Global Games Market Will Generate \$152.1 Billion in 2019 as the U.S. Overtakes China as the Biggest Market. *Newzoo*. <https://newzoo.com/insights/articles/the-global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/>
- Yang, B., Liu, Y., Liang, Y., & Tang, M. (2019). Exploiting user experience from online customer reviews for product design. *International Journal of Information Management*, 46, 173–186. doi:10.1016/j.ijinfomgt.2018.12.006
- Zhou, Y., & Yang, S. (2019). Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews. *IEEE Access: Practical Innovations, Open Solutions*, 7, 27769–27780. doi:10.1109/ACCESS.2019.2901472

Zhi Wang graduated from MSc in Business Analytics of Xi'an Jiaotong-Liverpool University, Suzhou, China. Victor Chang (Prof.) is a Professor of Data Science and IS at Teesside University, UK. He was a Senior Associate Professor, Xi'an Jiaotong-Liverpool University between June 2016 and Aug 2019. He was a Senior Lecturer at Leeds Beckett University, UK between Sep 2012 and May 2016. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (HE, Fellow, Greenwich) while working for several projects. Before becoming an academic, he achieved 97% on average in 27 IT certifications. He won an IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 & 2018, 2016 European award: Best Project in Research, 2017 Outstanding Young Scientist and numerous awards since 2012. He is widely regarded as a leading expert on Big Data/Cloud/IoT/security. He is a visiting scholar/PhD examiner at several universities, an Editor-in-Chief of IJOCL & OJBD, former Editor of FGCS, Associate Editor of TII & Info Fusion, founding chair of international workshops and founding Conference Chair of IoTBDS, COMPLEXIS, FEMIB & IoTBDS. He was involved in projects worth more than £13 million in Europe and Asia. He published 3 books and edited 2 books. He gave 22 keynotes internationally as a top researcher.

Gergely Horvath (PhD) is an Assistant Professor of Economics at the Division of Social Sciences of the Duke Kunshan University (DKU), a Sino-US liberal arts college founded by Duke University and Wuhan University as a joint venture. DKU is located in Kunshan, Jiangsu Province, China, near to Shanghai and Suzhou. He received a Ph.D. in Economics from the University of Alicante, Spain in 2011. Previously, he worked as an Assistant Professor at the School of Public Administration of the Southwestern University of Finance and Economics (Chengdu, China) between 2011 and 2014 and as a Postdoctoral Researcher at the Chair of Economic Theory (Prof. Dr. Veronika Grimm) of the Friedrich-Alexander University of Erlangen-Nuremberg between 2014 and 2016. Subsequently, he was a Lecturer in Economics at the International Business School of the Xi'an Jiaotong Liverpool University between 2016 and 2020, where he served as the Programme Director of the MSc Business Analytics program.