Text Analytics Project

Analyzing Key Features to Improve Game Experience using Steam Reviews

By,

Hariharan Murali Krishnan

Yagna Venkitasamy

Haritha Diraneyya

Naman Mehta

Executive Summary

There has been a steady increase in the popularity of computer games which has made it a multi-billion dollar industry. The game development industry is highly competitive which makes developing a successful game challenging. In addition, prior studies show that gamers are extremely hard to please, making the quality of games an important issue. Most online game stores allow users to review a game that they bought. Such reviews can make or break a game, as other potential buyers often base their purchasing decisions on the reviews of a game. Hence, studying game reviews can help game developers better understand user concerns, and further improve the user-perceived quality of games.

This report aims to identify key features that determine the gaming experience of an adventure game and quantify its effect on the likelihood of a game being recommended by reviewers.
To identify the important factors, topic modelling analysis was done on the game reviews for 10 adventure games which covered different subtypes of adventure games. Since different games have different Pros and Cons that influence the recommendation of the game, we have used the factors extracted from topic modelling to analyze the marginal effects of these features on game recommendation for a game. For this analysis, “Citadel: Forged with fire” game reviews, scraped from Steam website is used.

Key features identified from the analysis are: Game play, sound design, ambience, story, animation, humor, puzzle, character, difficulty. Our analysis suggests that for “Citadel: Forged with fire” game, when the review talks about the ambience and sound design of the game, the odds of the game being recommended increases by 219% and 176% respectively, which suggests that ambience is a strong positive aspect of the game. Also when the review talks about the animation and puzzle design of the game, the odds of the game being recommended decreases by 75% and 26% respectively which suggests that animation is a strong negative aspect of the game with scope for improvement.

These findings suggest that the developers did a good job in sound design and ambience, but poorly in animation and puzzle design. This can help developers steer their efforts in the right direction, including further research. The results from the model corroborated with the fact there was a major complaint with regards to a disconnect of visual effects (animation) and actual physics engine effect area, which can be so annoying to players.

Table of Contents

1. Problem Significance

In recent years, the video game industry has silently taken over the entertainment industry. Today, its size significantly surpasses the film and music industries combined! It was natural that such a huge industry attracted much recent efforts, especially in the case of the leading studios that create blockbusters. However, even in this time and age where the video games industry has become massive, a closer look would show us that industry size is significantly different according to genre. A good example of this would point-and-click adventure games, a genre known for its focus on story and writing and puzzle design on top of anything else. While it had its glory in the past when the video game industry as a whole was relatively small, nowadays it shrunk to small pockets of specialized community fans, away from the mainstream gaming crowds. It also happened that the nature of this genre has allowed its games to be efficiently created by small indie teams with tiny budgets, as opposed to being produced by mega studios. As a result, this genre of video games has received far less attention in terms of research than the more common ones (e.g. First Person Shooting or RPG). In the absence of scientific research, developers often had to rely on their personal experience when looking for answers to their questions. One example of such questions pertains to the features of an adventure game. Since developers have limited resources (which better applies to indie developers in particular), data driven decision making is required to fully utilize the limited resources to design different features of a game. In this report we analyze how a particular feature (for ex., graphics) would affect the odds of an adventure game being recommended.

Below is a list of the game features we chose to analyze which we think are the important features to look for in an adventure game based on our research looking at various games:

- Good writing/story/dialogue
- Graphics/artwork/animation
- Voice acting; ambience/music/SFX
- Controls; good game length (usually that means 5 - 10 hours of gameplay in this genre)
- Humor/funny/good jokes
- Good puzzle design
- Interesting/deep characters.

This study aims to see what features have the most significant impact on the probability of a game being recommended by users. Equipped with such information, indie developers should be in a better position to distribute their limited resources in the most efficient way possible and help them make better development decisions.

2. Prior Literature Review:

We reviewed many literature that were done before but we were able to gain considerable insights from the below mentioned works:

1. **An Empirical Study of Game Reviews on the Steam Platform :**

In this paper, they performed an empirical study on the reviews of 6,224 games on the Steam platform. They studied the number and the complexity of reviews, the type of information that is provided in the reviews, and the number of playing hours before posting a review. This study helped us to understand the various text analytics methods that can be deployed for feature extraction from the reviews.

2. **Topic modeling and sentiment analysis to pinpoint the perfect doctor :**

In this fun project by one PhD candidate, they have created a dashboard of doctor snapshots using the Yelp reviews. They have made feature selection using the LDA and word2vec models to identify the key features that would contribute to the positive or negative sentiments among the patients that would impact the review of the professional. The work took in 187 doctors and 5880 reviews to run their models.

3. **Predicting the helpfulness of game reviews:** A case study on the Steam store:

This study evaluates the helpfulness of game reviews on the online Steam store. It collects a large set of user reviews of different game genres and builds a classification model to predict whether these reviews are helpful or not. This model can accurately predict the helpfulness of the reviews based on different thresholds. This work also investigates various types of textual and word embedding features and analyzes their importance for predictions. Furthermore, it develops a regression-based model that can predict the score or rating of game reviews on Steam.

4. **Video Game Experience Prediction**:

This study aims to correlate the gamer's experience with the game as the number of hours played and the review that he'd post in the steam website with the number of hours he played on Steam. This study helped us in identifying the web scraping methods to build our corpus.

3. Data Sourcing/Preparation

Games these days, even those published by big studio names, do sell their games through big publishing platforms. Steam is the largest publishing platform for PC games in the world. It is hard not to find an important title on Steam today. Since genuine reviews on Steam can only be written by verified purchasers, Steam reviews are generally considered to be reliable and reasonably trustworthy (especially when they align and seem consistent with reviews written by editorial websites that specialize in critiquing games). In a Steam review, each user either recommends a game or does not. Therefore, our "Y" variable in this study is a binary variable which tells if the game has been recommended or not by the reviewer.
Some criteria were taken into consideration when choosing the games for our study:

> **Genre**. It was ensured that the selected games belonged to the same genre so that the findings are more useful across those different games. But we also wanted to find out whether the findings are noticeably different when applying the same analysis to these different games.

The data was collected from the reliable source of Steam website reviews. We built the corpus by web scraping the data from the website using API call.

Variables	Column Names	Description	Source
Game ID	game	Unique identifier of the game	Steam
Recommended	voted_up	Binary variable if the game has been recommended by the reviewer	Steam
Review text	review	Actual text review given by the reviewer	Steam

Table 2.1

The dataset we got had 10 games and 13000 reviews with 16 data attributes. It was based on gamers' feedback about the game he played. The dataset was in JSON files which required extensive data cleaning and formatting. We worked on the data to remove data inconsistencies

and missing values to clean the data as this would impact the performance of the model. Many of the observations had missing values

Next comes the text cleaning part. We first used a custom filter to filter out the stop words, curse words, punctuations, numeric data, emojis. We then used parts-of-speech tagged lemmatization. This drastically reduced the number of words in corpus when compared to the normal lemmatization technique. We used this corpus to train a word2vec model for vectorizing the corpus. This model was then used to calculate cosine similarity metric later in the analysis.

For testing the features, we selected the game, “Citadel: Forged with fire” that contained mixed reviews. The corpus was cleaned in a similar method and the corpus was used for analyzing the game.

4. Exploratory Data Analysis/ Visualization

Data was first explored, not only to get a general idea of the data being dealt with, but also as part of the topic extraction that was done later. Now that the text has been arranged in Word2Vec, it became possible to apply sentiment analysis. POS filtering was applied. By lemmatizing the words, the insights regarding word frequency became more meaningful. The word “game” was the most frequent word (i.e. post-applying the POS filter), at least seven times more frequent than the next frequent word “time”, used more than 7000 time across the reviews. However, using weighted frequency for topic extraction, we came with a list of key features of the reviewed game, the most significant being: humor, ambience, sound, puzzle, and animation.

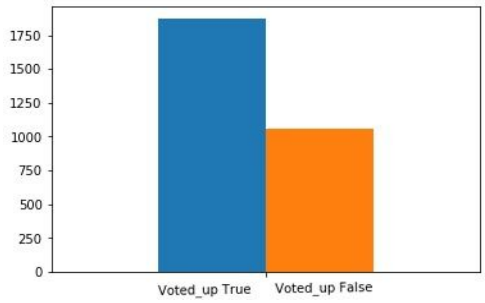
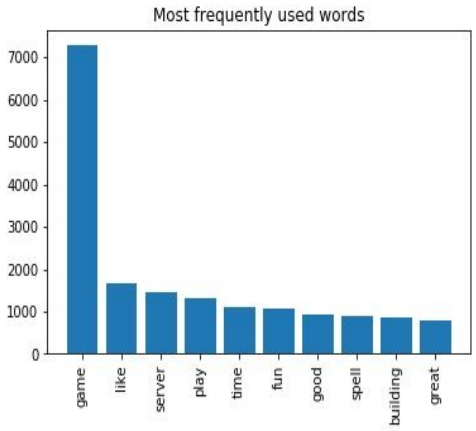
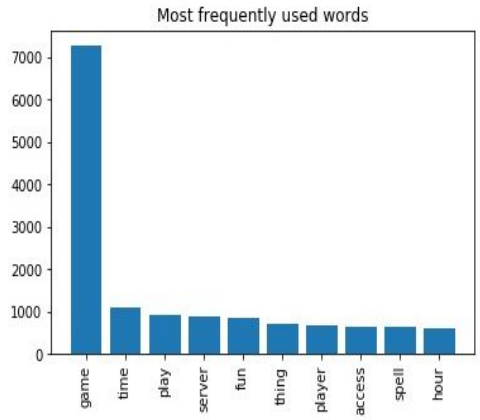


Figure 3.1

Figure 3.1 displays the spread of the dependent variable between the two classes. We can see that our data is imbalanced with one class having more “voted up” true than “voted up” false.



Without POS Filter- Figure 3.2.1



With POS as Noun only - Figure 3.2.2

Figure 3.2.1 displays the most frequently used words in the corpus without the parts of speech filter. We can see that unimportant words “like”, “good”, “great”, which are listed here which are not much useful in our analysis part. Hence we implemented the parts of the speech filter with only nouns that gave us good features to concentrate on in Figure 3.2.2.

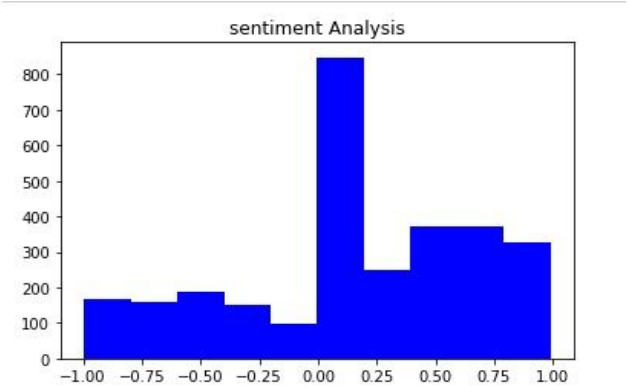


Figure 3.3

Figure 3.3 displays the Sentiment Analysis. This is consistent with the fact that positive reviews outnumbered negative ones by about 70%.

5. Text Analysis and results

Key topic extraction:

For extracting the key features that are important for an adventure game, we used topic modelling technique. The corpus for the topic model consisted of a corpus built from the 10 games which cover different subtypes of the adventure game genre. Different LSI and LDA models with different topic counts were tested to find topics that made sense. The LSI model trained on Noun filtered POS tagging gave us good results and the key words in the topics were used as features for analyzing the reviews.

Feature	Description
ambience_sim	Does it give the scenes the right “feel”?
sound_design _sim	Is the game world rich with sound FX?
humor_sim	How is the quality of the game’s humor?
story_sim	The quality of the story: is it hooking and well written in terms of interesting characters and building climax?
gamplay_sim	Is the gameplay experience smooth? Is it on the other hand cumbersome? Are there factors impeding that experience such as poorly designed user interface?
difficult_sim	Does the game have the right level of difficulty? Ideally it should not be too difficult to the level where it is discouraging, but also not too easy as not to be challenging.
character_sim	Are the characters interesting and are they aesthetically appealing?
puzzle_sim	How well are the puzzles designed? Does their logic make sense? Is their level of difficulty just right (not too difficult or too easy)?
animation_sim	3D design, animation of characters and moving objects, visual effects.

Vectorizing the data:

For vectorizing the data, the bag of words approach was not appropriate because a simple frequency distribution of the reviews wouldn’t capture different words used to comment on the

same feature. For ex., humour can be expressed with different words like comedy, hilarious etc. Hence a word2vec model trained with 10 games reviews was used for vectorizing the data.

Feature extraction using Cosine Similarity distance:

The keywords from the topic models along with other features from prior research were used as independent variables for analyzing the patterns in recommendation. A custom function was defined to calculate the similarity of the feature (say graphic) with each word of the review for those words which are above a certain threshold similarity level to filter out words which are unrelated to the feature. The similarity score has a range [0,1].For ex., when sound_design has "0"score, it implies that the review doesn't talk about sound_design. A value of 1 suggests that the review mainly focuses on the sound_design aspect. Since a particular review can talk about multiple aspects of the game, different scores are assigned to each column for a review. This ensures that all the content in the review is captured.

clean_words	review	voted_up	gameplay_sim	sound_design_sim	ambience_sim	story_sim	animation_sim	humor_sim	puzzle_sim	character_sim	difficult_sim
[yes]	Yes	True	0.601	0.0	0.597	0.0	0.0	0.000	0.0	0.0	0.556
[cool, game, ark, rust, exist]	It would a cool game if Ark, Rust, etc didn't ...	False	0.514	0.0	0.587	0.0	0.0	0.000	0.0	0.0	0.000
[friend, grabbed, look, new, play, man, surpri...	Me and a few friends grabbed this looking for ...	True	0.000	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
[havent, played, wonderful,	I havent played much of this	True	0.000	0.0	0.630	0.0	0.0	0.000	0.0	0.0	0.582

Models : (Logistic regression and Random Forest)

The features extracted using the above technique was used to train a model, a logistic regression model and a random forest model. The random forest ensemble model provides a feature importance metric to understand how the models predict the recommendations. This value is calculated based on the purity of the nodes generated from the split criteria in different trees. To understand the marginal effects of the features on recommendation we ran a logistic regression model which shows how the odds of game recommendation changes for different aspects/features of the games.

The random forest slightly had a better f1-score when compared to random forest. But, we chose the logistic regression model to explain our results because of model explainability.

Logit Model - Results

Features	Coeff	Std err	exp of coeff	Marginal effect of odds
ambience_sim	1.1601	0.1687	3.190252285	219%
sound_design_sim	1.0164	0.3336	2.763229212	176%
humor_sim	0.9718	0.2972	2.642697035	164%
story_sim	0.1954	0.2285	1.215797208	22%
gameplay_sim	-0.1137	0.1865	0.892525673	-11%
difficult_sim	-0.1265	0.1835	0.88117415	-12%
character_sim	-0.1605	0.4308	0.851717824	-15%
puzzle_sim	-0.2945	0.3766	0.744903946	-26%
animation_sim	-1.4063	0.229	0.245048287	-75%

1. When the review solely talks about ambience, the odds of the game being recommended increases by 219% which suggests that ambience is a strong positive aspect of the game according to the reviewers.
2. When the review solely talks about animation, the odds of the game being recommended decreases by 75% which suggests that animation is a strong negative aspect of the game according to the reviewers

Random Forest - Results

	importance
ambience_sim	0.314027
gameplay_sim	0.165782
difficult_sim	0.152790
animation_sim	0.096124
story_sim	0.090160
humor_sim	0.072856
sound_design_sim	0.057130
puzzle_sim	0.031529
character_sim	0.019602

The results of the random forest also paints a similar picture which suggests that ambience is a strong feature in predicting the recommendation of the game followed by gameplay. This tells us that the results from the logistic regression model are stable and trustworthy.

6. Quality Check

Once we identified the most significant independent variables, we attempted to check the reviews focusing on these identified variables to see if we can gain deeper and more specific insights that might be useful to the game developers. Here are some of our findings below:

Related Feature	**Specific Issue**	**Details**	**Fixable by Game Patch?**
Animation	Disconnect between Visual Effects and the Physical Engine Area of Effect	In action adventure games, many character abilities and world natural impacts affect an area, not a single character. Naturally these AOE (area of effect) abilities have animation and visual effects to let the player know the correct AOE to dodge if it is a negative (e.g. damaging) AOE or enter it if it is a beneficial AOE (e.g. healing). It is a bad thing if the actual AEO as defined by the game's physical engine and the visual effect do not match as the player won't be able to correctly respond to the AEO effect. This is an "animation" related issue that we found a few reviews talking about.	This shortcoming can be easily fixed by a game patch update and does not need a fresh game release.
Gameplay	Bad UI (User Interface)	Badly designed UI can significantly impede the play experience. Shuffling through items is one of the most basic game mechanics in an adventure game (considered part of the UC), and found the UI of the game Citadel to be Strongly criticized.	Same as above.
Difficulty	Game Became More Difficult after an Update	Some reviews were found complaining that the game difficulty level used to be much better before a recent patch update.	Same as above.

7. Insights

Based on our statistical analysis, we can derive the following insights from our data:

Key features identified from the analysis are: Game play, sound design, ambience, story, animation, humor, puzzle, character, difficulty.

Our analysis suggests that for “Citadel: Forged with fire” game, when the review talks about the ambience and sound design of the game, the odds of the game being recommended increases by 219% and 176% respectively, which suggests that ambience and sound design is a strong positive aspect of the game.

When the review talks about the animation and puzzle design of the game, the odds of the game being recommended decreases by 75% and 26% respectively which suggests that animation is a strong negative aspect of the game with scope for improvement.

Based on the results from the analysis the game developer can focus his/her efforts to improve animation and puzzle design aspect of the game in future updates as this could greatly enhance the gamer experience

References

Literature Review Links:

1. https://www.researchgate.net/publication/324923032_An_Empirical_Study_of_Game_Reviews_on_the_Steam_Platform
2. https://blog.insightdatascience.com/topic-modeling-and-sentiment-analysis-to-pinpoint-the-perfect-doctor-6a8fdd4a3904
3. https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179022
4. https://github.com/mulhod/reviewer_experience_prediction/blob/master/apln_paper/reviews.pdf

Appendix

Python Code:

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(random_state=42)

classification_model(lr, x2.drop(["voted_up","sentiment_polarity"], axis = 1), x2.voted_up)
import statsmodels.api as sm
logit_model=sm.Logit(x2.voted_up, x2.drop(["voted_up","sentiment_polarity"], axis = 1))
result=logit_model.fit()
print(result.summary2())
```

```
 Train data stats:
[[ 136  916]
 [ 115 1745]]
              precision    recall  f1-score   support

           0       0.54      0.13      0.21      1052
           1       0.66      0.94      0.77      1860
                                .
    accuracy                           0.65      2912
   macro avg       0.60      0.53      0.49      2912
weighted avg       0.61      0.65      0.57      2912

CV Accuracy (5-fold): [0.6483705  0.61921098 0.64089347 0.6580756  0.64089347]
Mean CV Accuracy: 0.6414888036167943
Optimization terminated successfully.
         Current function value: 0.623735
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.046
Dependent Variable: voted_up         AIC:              3650.6352
Date:               2020-04-23 23:26 BIC:              3704.4246
No. Observations:   2912             Log-Likelihood:   -1816.3
Df Model:           8                LL-Null:          -1904.9
Df Residuals:       2903             LLR p-value:      4.2091e-34
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
-----------------------------------------------------------------
                  Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
-----------------------------------------------------------------
gameplay_sim      -0.1137   0.1865 -0.6098 0.5420 -0.4792  0.2518
sound_design_sim   1.0164   0.3336  3.0467 0.0023  0.3625  1.6702
ambience_sim       1.1601   0.1687  6.8754 0.0000  0.8294  1.4908
story_sim          0.1954   0.2285  0.8553 0.3924 -0.2524  0.6433
animation_sim     -1.4063   0.2290 -6.1401 0.0000 -1.8551 -0.9574
humor_sim          0.9718   0.2972  3.2694 0.0011  0.3892  1.5544
puzzle_sim        -0.2945   0.3766 -0.7820 0.4342 -1.0326  0.4436
character_sim     -0.1605   0.4308 -0.3725 0.7095 -1.0048  0.6838
difficult_sim     -0.1265   0.1835 -0.6891 0.4908 -0.4862  0.2333
=================================================================
```

```
from sklearn.ensemble import RandomForestClassifier
rfc2 = RandomForestClassifier(n_estimators= 100)
classification_model(rfc2, x2.drop(["voted_up","sentiment_polarity"], axis = 1), x2.voted_up)
#rfc2.feature_importances_
feature_importances = pd.DataFrame(rfc2.feature_importances_,
                                   index = x2.drop(["voted_up","sentiment_polarity"], axis = 1).columns,
                                    columns=['importance']).sort_values('importance', ascending=False)
feature_importances
```

```
 Train data stats:
[[ 383  669]
 [  82 1778]]
              precision    recall  f1-score   support

           0       0.82      0.36      0.50      1052
           1       0.73      0.96      0.83      1860

    accuracy                           0.74      2912
   macro avg       0.78      0.66      0.67      2912
weighted avg       0.76      0.74      0.71      2912

CV Accuracy (5-fold): [0.67409949 0.62778731 0.65635739 0.68041237 0.68900344]
Mean CV Accuracy: 0.6655319976658237
```

]:

	importance
ambience_sim	0.314027
gameplay_sim	0.165782
difficult_sim	0.152790
animation_sim	0.096124
story_sim	0.090160
humor_sim	0.072856
sound_design_sim	0.057130
puzzle_sim	0.031529
character_sim	0.019602