Text Classification of Steam User Reviews

Classifying Reviews Using Steam Data Set

Executive Summary

- Using 9.6 million English reviews from the Steam Library, we created a model that can classify user reviews as positive or negative with 91% accuracy. The model can help classify reviews for new games in development to find consumer sentiment and feedback.

Introduction

- Steam is a leading distributor of digital games, with an estimated annual revenue of 8.88 billion in 2022.[1]
- Many game developers generate a majority of their revenue through Steam.
- There is an estimated 132 million monthly active players in 2021 on the Steam, making steam one of the top PC gaming platforms available worldwide.
- Steam has a vast variety and sample of games to conduct research based on reviews provided from users.
- Through the use of Steam library user reviews, it is possible to determine player sentiment for games in early development and testing stages.

1. Clement, J., “Steam Gaming Platform - Statistics & Facts,” Statista, https://www.statista.com/topics/4282/steam/

Data Features

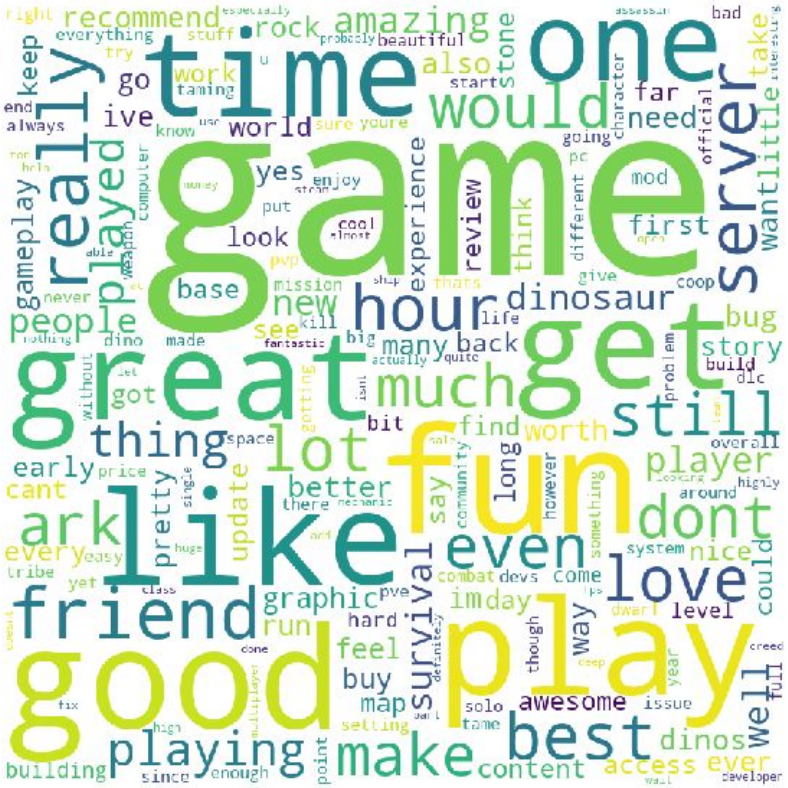
- The Steam dataset consists of reviews in multiple languages, in English there are over 9 million reviews.
- Each review contains different features such as:
  - Game recommended or not
  - Hours user played prior to publishing review
  - Steam purchase
  - Received for free
  - Various user information, such as number of games owned.

Data Wrangling

- User inputs without written reviews were removed.
- A portion of the data was sampled for our model, 263,815 values were used selected at random 30% of the games available.
- Each review was then converted into tokens to separate each review into meaningful units while removing excess information such as stopwords (ex. the, and, then.)

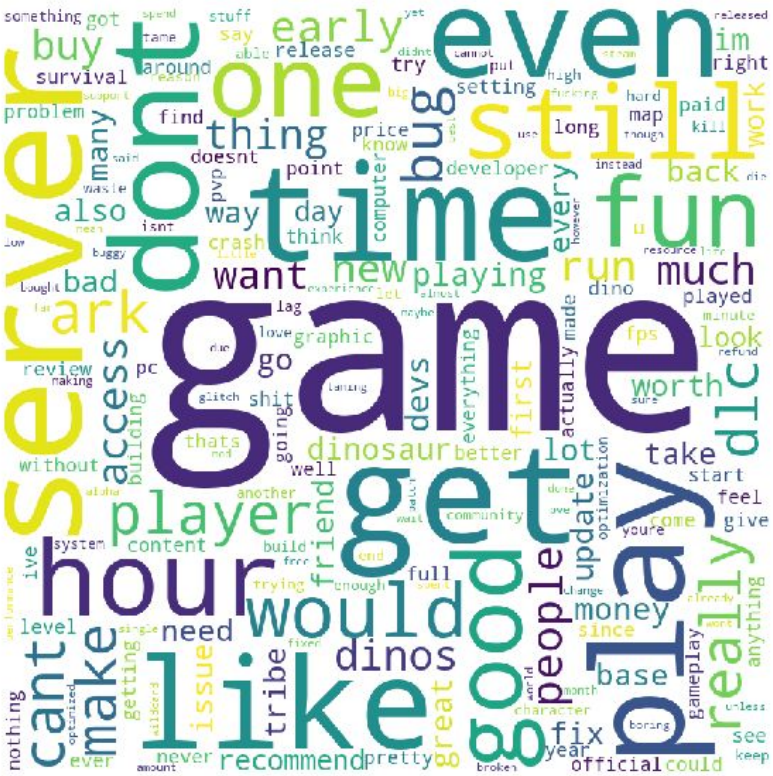
Data Summary-Positive Words

- Word Cloud depicting the frequency of all words in positive reviews.
- Positive bigrams and trigrams which have larger effect on our model include:
  - (great, game)
  - (good, game)
  - (best, game, ever)



Data Summary-Negative Words

- Word Cloud depicting the frequency of all words in negative reviews.
- Negative bigrams and trigrams which have larger effect on our model include:
    - (early, access)
    - (play, game)
    - (paid, dlc)
    - (early, access, game)



Early Data Analysis - Hypothesis Testing

Hypothesis: The proportion of games recommended during early access is different than the proportion of games recommended after release date.

H_0: $p_{ea} = p_{rd}$

H_a: $p_{ea} \neq p_{rd}$

Our p-value < α at 1% significance level, we reject the null hypothesis H_0 there is a difference between game recommendations in early access and launch.

Early Data Analysis - Hypothesis Testing

Hypothesis: The proportion of games recommended given for free is different than the proportion of games recommended paid by consumers.

H_0: $p_{free} = p_{paid}$

H_a: $p_{free} \neq p_{paid}$

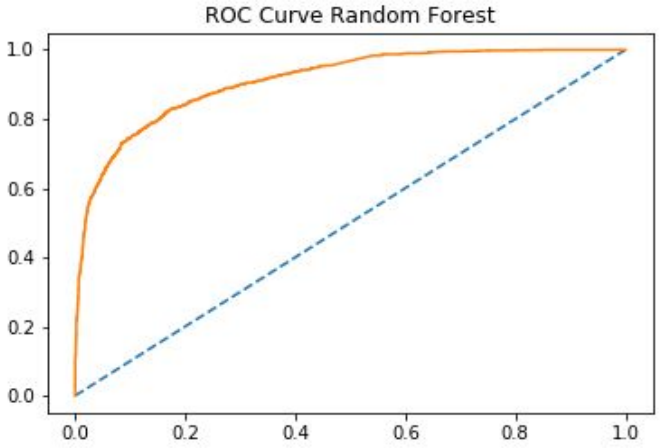
Our p-value < α at 1% significance level, we reject the null hypothesis H_0 there is a difference between game recommendations of gifted games and purchased games.

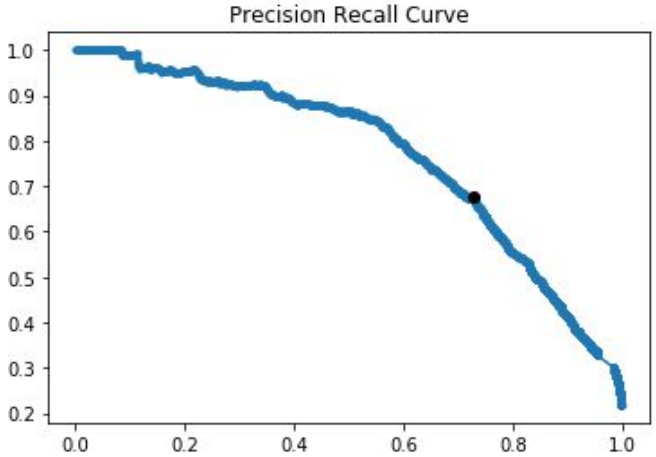
Machine Learning Model

- After running a gridsearchCV using two different models (Multinomial Naive Bayes and Random Forest Classifier) the random forest classifier performed best.
- Then using our optimized model to classify our test set had an AUC score of 0.911, meaning the model has above 90% chance of classifying a review sentiment correctly.

Results

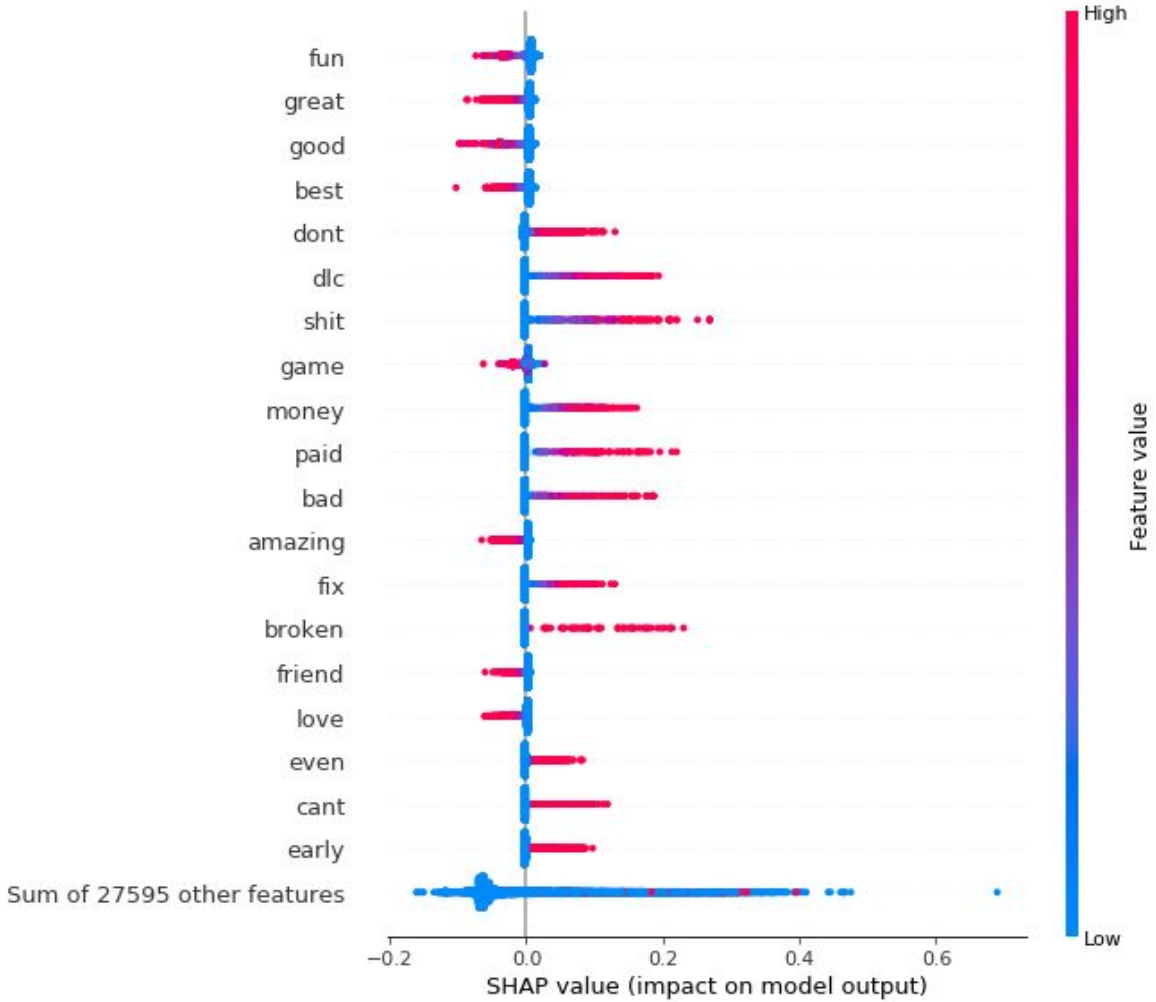
- The ROC curve indicates that the Random Forest model does contain skill above the dotted baseline.
- The precision recall curve also depicts that our model has high precision and high recall, but we can also adjust accordingly to our needs, which is the dot on our graph.
- The model is optimized to provide high true positives while minimizing false negatives.





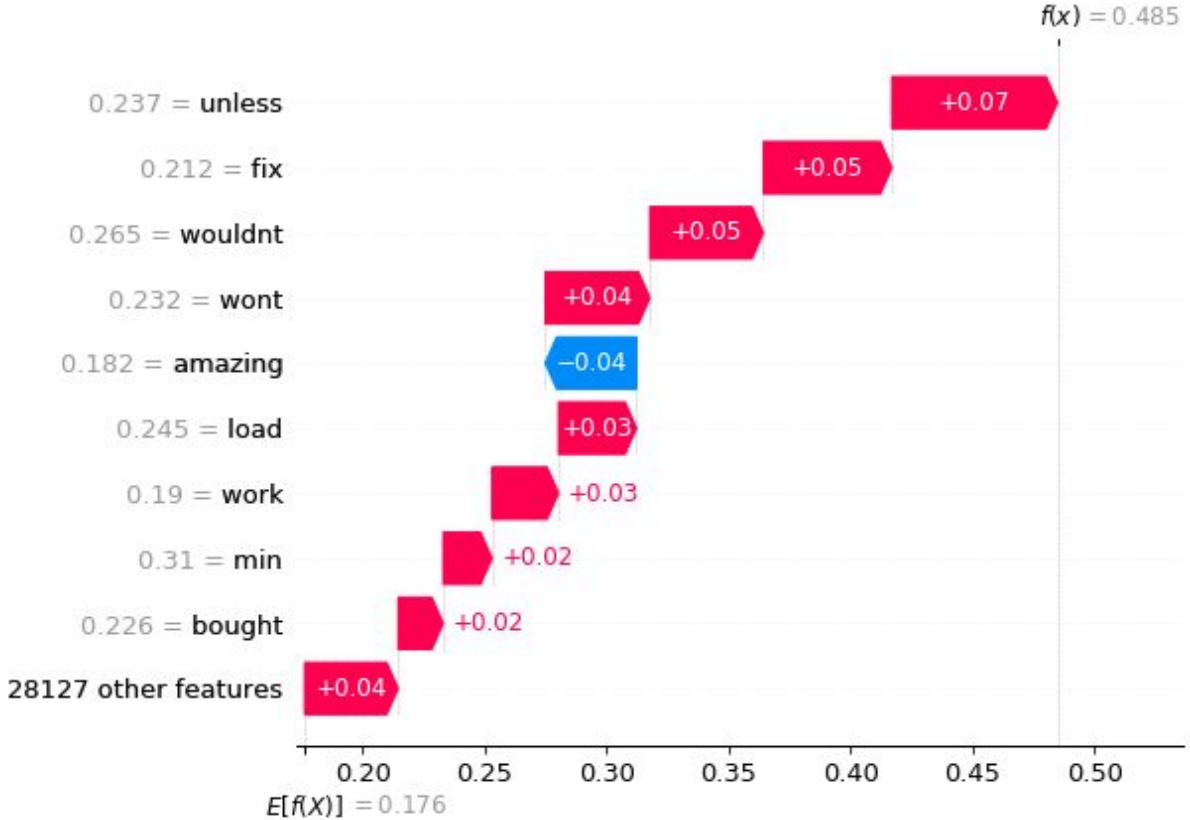
Beeswarm SHAP Plot

- We used SHAP values to visually depict how features affect our model.
- From the table we can see the words in reviews that highly impact the model outputs.
- Most negative reviews contain ‘DLC’, ‘paid’, and ‘broken’. Errors which game developers can look into to adjust or fix.
- These types of plots can bring insight on what impacts their consumers reviews.



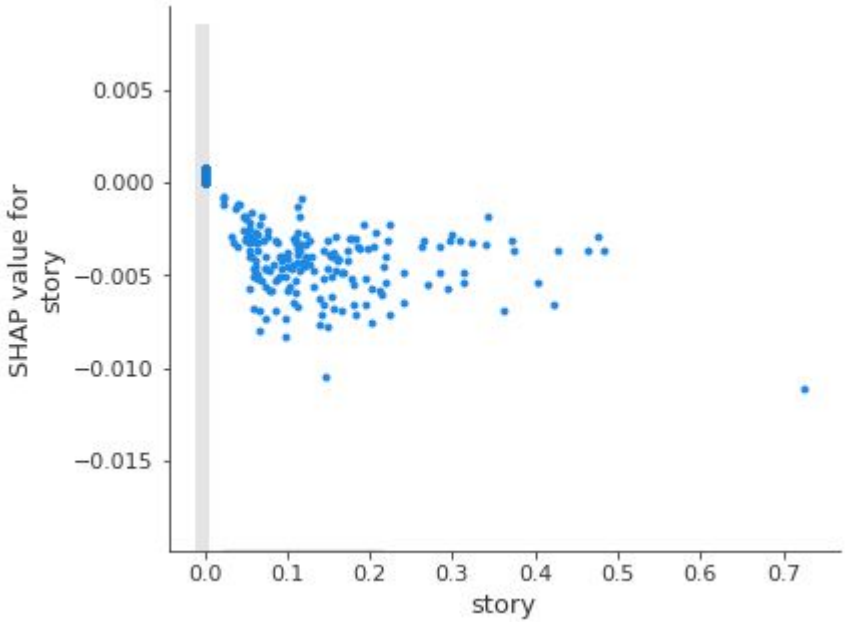
Classification Example

- A consumer wrote the review: “bought game never load computer supasses min spec still wont work seems like amazing concept could ever play wouldnt recommend unless fix problem”
- Our model predicted the review sentiment is negative.
- From our SHAP waterfall plot we can see how each feature contributed to the outcome of our model visually.



Variable Sentiment Search

- Can isolate variables in their own SHAP scatter plots to see their impact on the model.
- For example, if the variable “story” is of importance we can search reviews to see the impact this variable has on our model.
- The scatter plot trends downward meaning each time the variable “story” is present the prediction leans towards a positive sentiment.
- Also worth noting: When the variable “story” is zero or not present, there is a small shift toward a negative review sentiment.



Application and Further Research

- Our model can help game developers classify their consumer sentiments to provide feedback on their games. Feedback and testing is important to ensure they can produce an optimized game to maximize their profits.
- The next step in the analysis could include theming where categories are created based on areas of concern. For example, pricing, graphics, sound and etc. How consumers view these categories could help developers further improve their games.