Text Classification of Steam User Reviews

Classifying Reviews Using Steam Data Set

Executive Summary

- Steam library is one of the largest PC digital game distributors in the market.
- Steam reviews data set consists of 9.6 million English reviews of games ranging from triple A games to smaller indie games.
- Each review contains the user opinion, positive or negative sentiment and number of users who found the opinion helpful. After data wrangling, each review contains tokenized text of the user generated reviews.
- The best model created was a Random Forest Classifier, which is able to predict a sentiment score per user review. This model can be used to evaluate newly developed game reviews to predict user sentiments.
- Our model has an AUC of 0.911, which is a 91.1% probability of classifying our user reviews correctly.

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.
- Therefore, 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

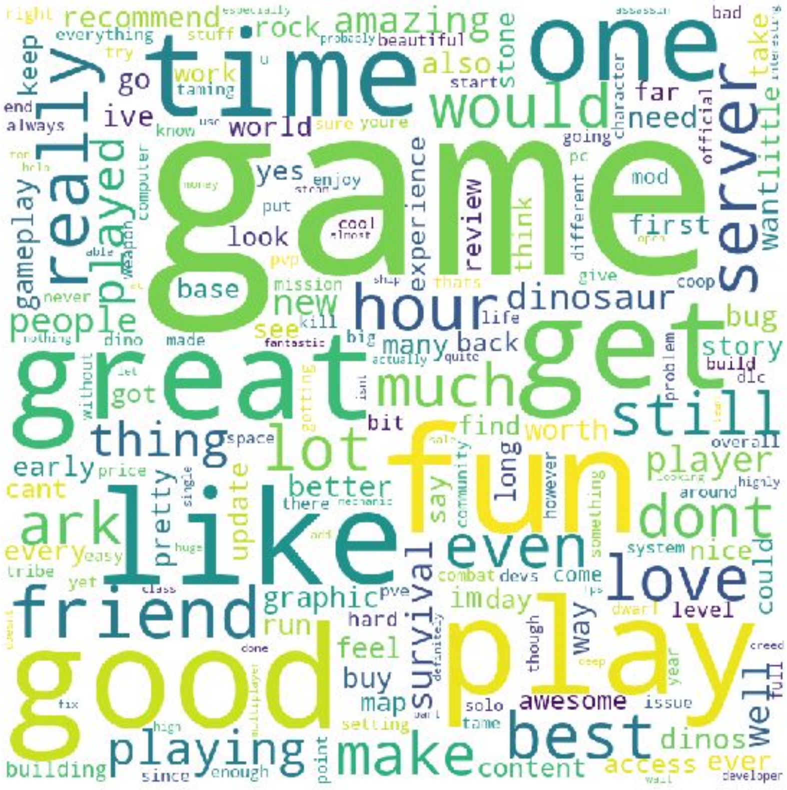
- The Steam dataset consists of reviews in multiple languages, in English there are over 9 million reviews.
- Each review contains different features such as:
  - Game recommended or not
  - Hours played prior to publishing review
  - Steam purchase
  - Received for free
  - Various user information,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

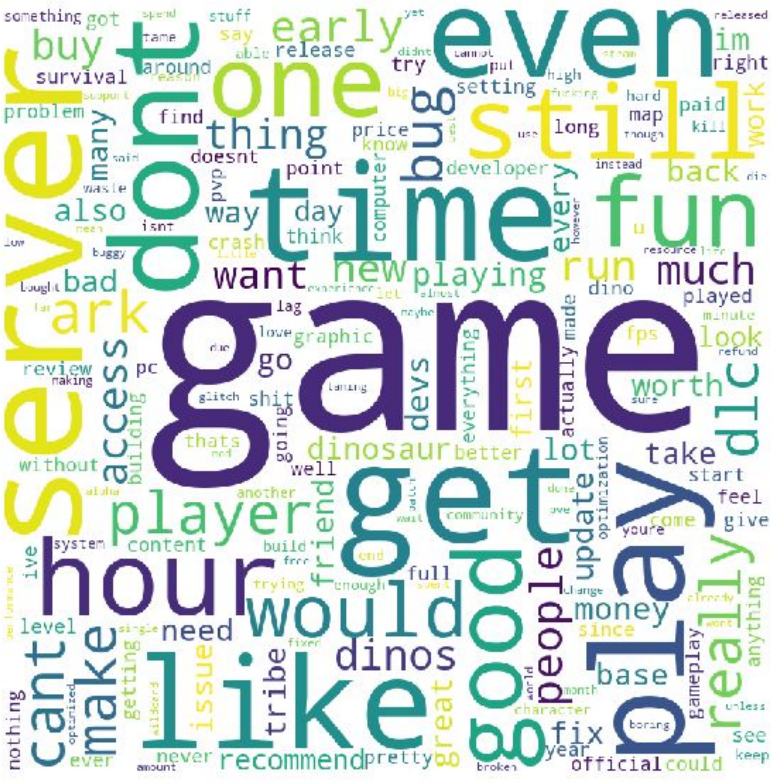
Data Summary-Positive Words

- Word Cloud of all positive words.
- Feature importance includes:
  - (great, game)
  - (good, game)
  - (best, game, ever)



Data Summary-Negative Words

- Word Cloud of all negative words.
- Feature importance includes:
  - (early, access)
  - (play, game)
  - (paid, dlc)
  - (early, access, game)

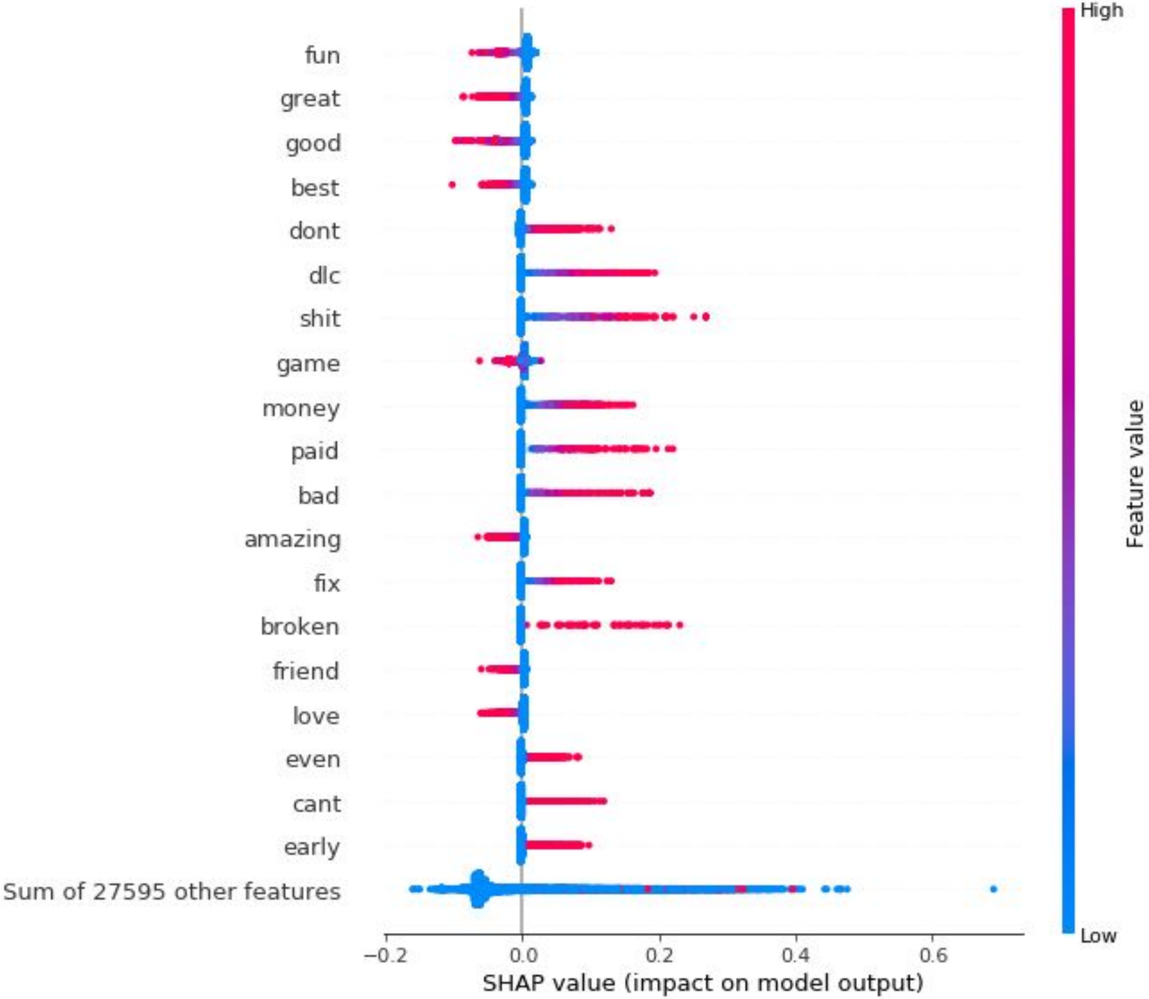


Machine Learning Model

- After running a gridsearch using two different models (Multinomial Naive Bayes and Random Forest Classifier) the random forest classifier performed best.
- The using the 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.

Beeswarm SHAP Plot

- Feature importance of reviews.
- Positive values are for review words that have high impact on negative reviews, while negative value words have high impact on positive reviews.
-



Classification Example

Application and Further Research