Abstract

The video game industry is hugely popular and rapidly growing, generating over 179 billion US dollars in global revenue in 2020, an increase of 20% over the previous year. The Steam platform, a video game distribution service that is the largest of its kind, is used by over 120 million users every month and provides a selection of over 60 thousand games.

This dissertation used a dataset of over 10 million video game reviews, taken from Steam, to investigate the degree to which the text of a review can be used by machine learning models, such as BERT, to predict other features of the review. The particular review features that were investigated were the review polarity, the community-determined helpfulness of the review and the amount of time the reviewer spent playing the game before reviewing it.

This investigation was also expanded upon and used to develop a method to determine which reviewers are most representative of the overall population using a somewhat novel approach. The methodology underpinning this new approach, the assumptions made when implementing it and the drawbacks that it presented were discussed. Potential improvements and refinements, alongside potential commercial applications, were also suggested.

The results of the review feature prediction were mixed. The polarity of reviews was able to be accurately predicted, with accuracies reaching as high as 85.9% for balanced data samples and 91.3% for imbalanced samples. The models developed to predict review helpfulness did not prove successful, while the models developed to predict reviewer playtimes produced uncertain results. The models developed to determine representative users performed sufficiently better than baseline models, indicating that the approach, even in its simplest form, bears potential for further investigation. The model was then used to successfully identify users that it considered to be representative.