Video Game Reviews Text Analysis

By Fei Guo, Victor Li, Jonathan Lin, Xinyu Zheng

Abstract

Video game have reached a saturation point where there are devoted media sites to reviewing video games as well as news around the gaming industry. However, there has not been research done on game reviews the same way that other entertainment industries like movies and music have. We seek to find textual patterns and characteristics in video game reviews by scraping established video game sites and conducting latent dirichlet allocation, jargon distance, and sentiment analysis to identify common writer characteristics and habits. We found that most writers typically score games between 60 to 70 out of 100 and use similar phrases to describe those games in their reviews. Further, we found through sentiment analysis that most writers use positively associated words, even for games that score poorly. We suggest more detailed analysis for games in certain genres to see if these results are consistent within a subpopulation of the reviews and propose suggestions for future work which include a write recommendation system.

Introduction

The video game marketplace in North America has grown up tremendously from its birth in early 1980s. What was once a very small market competing for space with pinball machines in an arcade has grown to a multimillion dollar industry that employs tens of thousands of people to develop and create new games. This rise of a new entertainment industry coupled with the advent of the internet allowed for publications and websites of all sizes to begin covering the industry. This has resulted in a wealth of coverage on video games through previews and reviews.

Despite the increase in outlets following video games, the games journalism industry hasn't been long enough to establish notable critics the same way film and music has. This combined along with turnover at media outlets can make it hard to find similar reviewers that a person can identify with. With so many voices now writing and talking about both video games and the games industry, it can be overwhelming to find a writer who shares similar interests with a user.

The purpose of this project is to analyze textual similarities in reviews to connect certain reviewers together. This could be used to link writers with similar tastes and recommend other reviewers based on one reviewers scores or writing style.

Datasets Used

To analyze our question, we scraped three websites for game reviews: Giantbomb, IGN, and Gamespot. All three websites have reviews from employees who are paid to review games and stay current on news and trends in the video game industry. Gamespot and IGN were both founded in the mid 1990s and have been established websites about video games for over a decade. Both companies have a staff of about 25-30 editors on the content team. Giantbomb was formed together by previous editors at Gamespot in 2008 and currently employs 6 editors.

Reviews were scraped using a combination of the Kimono API, BeautifulSoup, and the Giantbomb API. The Kimono API allows for selections of elements on a webpage and extracts those elements and their respective URLs into a CSV formatted sheet. BeautifulSoup was then used to retrieve the text from those URLs and pull the full text reviews, scores, and authors associated for each of the game reviews. These were then placed into separate CSV formatted sheet. This method was used for both Gamespot and IGN, while for Giantbomb we used their API which let us grab the full text reviews, scores, authors all in one location. All game reviews were from the last 18 months with the exception of Giantbomb game reviews which date back to the 2008. The output were logged into a CSV formatted sheet as well.

All three CSV files from each website were merged together to create a combined CSV file containing all the full texts reviews and their associated metadata. Review scores from each website were converted over to a 100 point scale to standardize all the scores. We had 133 different writers, 1,431 distinct games reviewed, and 1,827 reviews. This would be the base dataset where we conducted our analysis.

Primary methods implemented and results

Initial Analysis

Our preliminary analysis began by first computing some basic statistics on the text reviews. We calculated the length of the reviews and averaged the lengths for each writer.This gave us a

sense of the average length that a paid writer usually needs for a review.

Figure 1: Average review length per reviewer

Most writers were around 1,250 words per review. Some of the more extreme values on the figure above are because of a small sample size for the writer. We then grouped the writers together by website to calculate average text review lengths for the three websites.

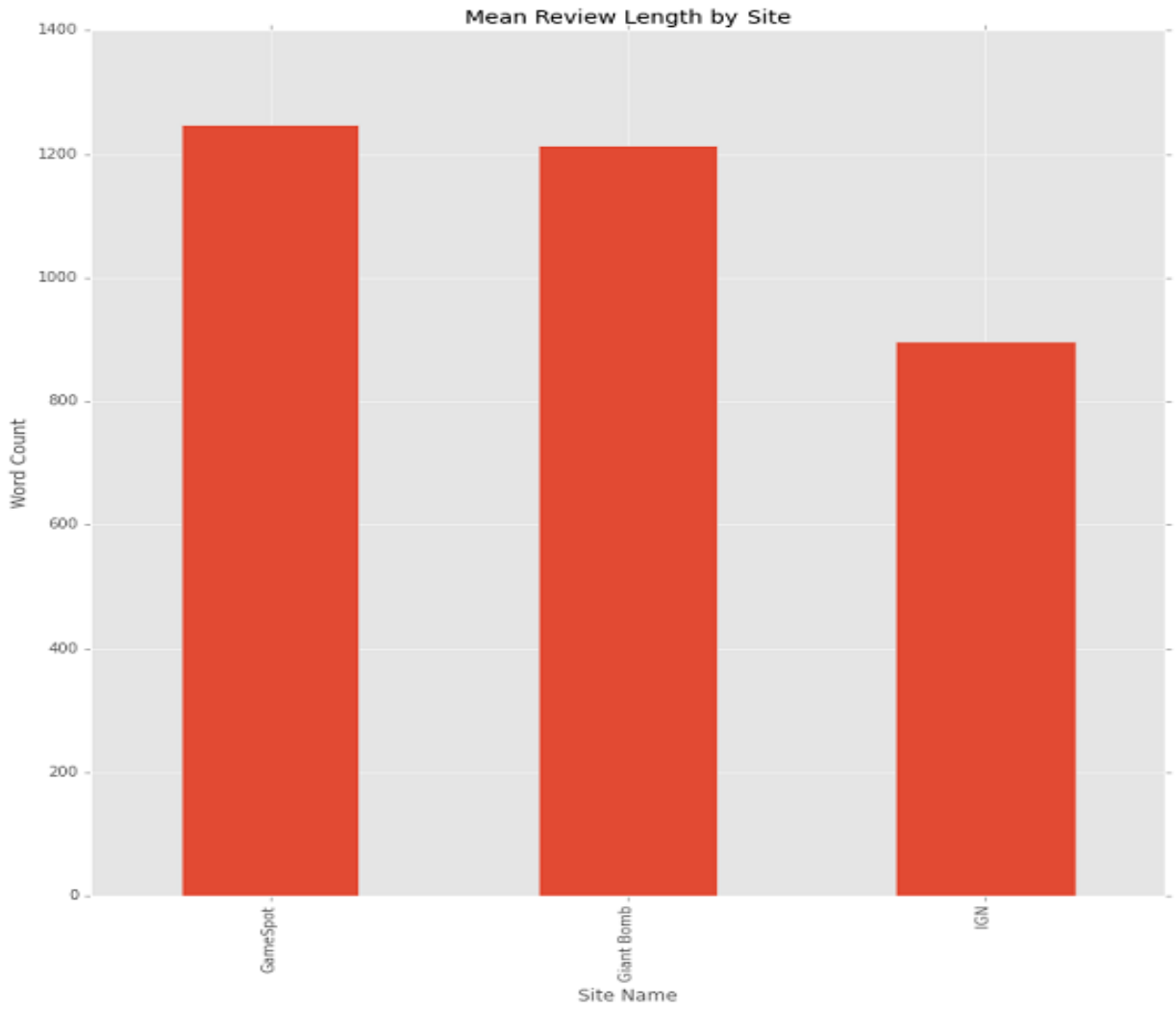


Figure 2: Average review length per website

IGN had the lowest average of the three websites, but they also cover more areas of entertainment such as movies. The lower word counts might be the result of IGN being unable to focus all of their energy into video games. Gamespot and Giantbomb are around the same word count which is interesting since Giantbomb has a much smaller editorial staff but has a lot of reviews in our dataset.

We then computed the same basic analysis on reviews scores, again for each reviewer and then for all reviewers.

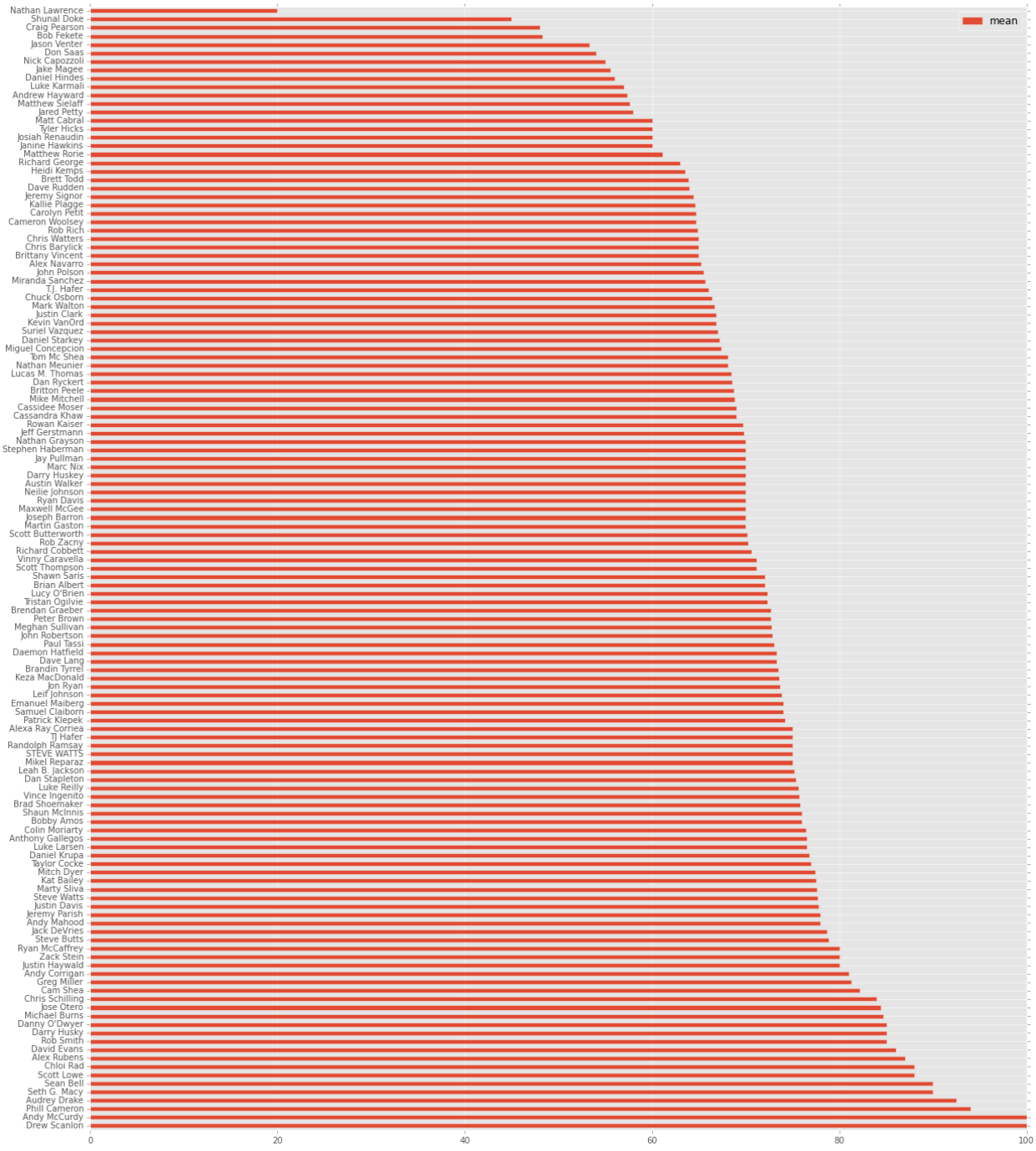


Figure 3: Average review score for each writer

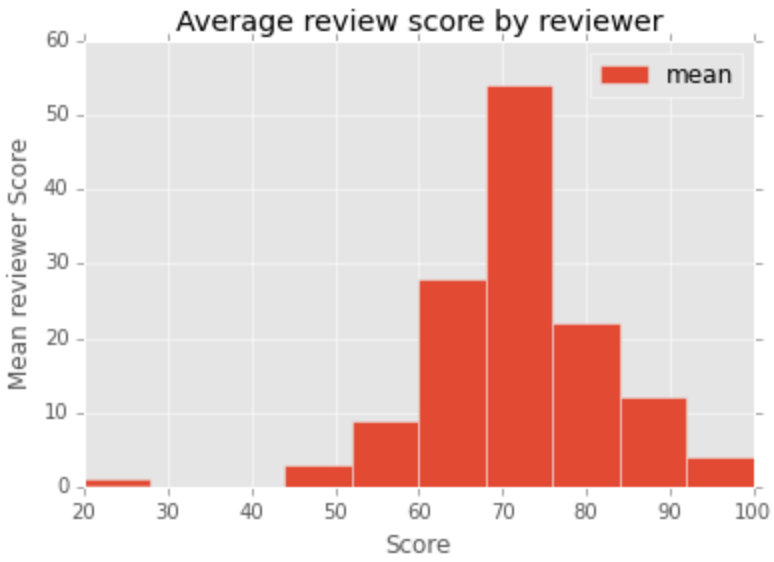


Figure 4: Average review score for all reviewers

These figures show that most reviewers score games around the 60-70 range after converting their scores over. Thus most games are considered to be good but not great by writers from these sites.

After conducting the preliminary analysis on the simple statistics, we began to focus on the actual text in the reviews for most insights.

Latent Dirichlet Allocation (LDA)

Topic modeling was employed to see if there are common terms used in reviews of lower rated video games. From our analysis above we selected 65 writers that had average review scores below the total average of all reviewers and analyze their text reviews using LDA. As a comparison, we also randomly pulled 65 writers and their text reviews to run the same analysis. The text was preprocessed using the Natural Language Toolkit python package for word tokenizing and elimination of stop words. Stemming was done in initial runs of the analysis but lead to unintelligible output, so we decided to not stem words after eliminating stopwords and other characters. Using the LDA package in python, we decided to find 15 topics for 1000 iterations on both sets. Listed below are the outputs from the LDA analysis:

For 65 randomly selected 'general reviews' (control group):

Topic 0: pursue hold splinter crimsonland explained replace nigh elaborate strike
Topic 1: increasing payoff lore locked elegant idarb reliable nigh tired
Topic 2: hordes bringing geography unit symphony spiders standout bonuses thunder
Topic 3: geography elegant four standout kids basics concepts reliable nigh
Topic 4: regional brings locked minigames concepts tired idarb reliable nigh
Topic 5: holy tired successful woods concepts blade spacebook standout idarb

Topic 6: increasing hero resilient bonuses thunder bolts valiant second elegant
Topic 7: electricity thunder military resilient hole valiant second elegant wooden
Topic 8: symphony guns-blazing golden r-type second wooden reliable valiant elegant
Topic 9: reliable music chassis irksome bolts concepts tired standout idarb
Topic 10: payoff tires hurt concepts elegant idarb reliable nigh tired
Topic 11: lord minigames spiders spacebook idarb valiant grueling r-type amzntargeting
Topic 12: increasing payoff holy crimsonland concepts tires idarb reliable nigh
Topic 13: unit brings lore reports blade brought standout bonuses looking
Topic 14: increasing holy opponents kids tired wooden standout idarb reliable

For 65 score-below-average reviews (treatment group):

Topic 0: elaborate splinter increasing meadows succumb woods minigames regional wooden
Topic 1: music second unit hole basics tweaked regional bringing starships
Topic 2: melvin spiders tweaked idarb brought minigames appropriately successful yellow
Topic 3: prices yellow hole wooden regional bringing starships basics scholar
Topic 4: valiant milestones four reports tires tired nigh succession tweaked
Topic 5: starships brought lore opponents tires hordes explained succession basics
Topic 6: nigh looking wooden hole minigames tires tired succession idarb
Topic 7: replace dna genre tired hole resilient glass basics four
Topic 8: danganronpa reports txk yellow four tired nigh succession idarb
Topic 9: nigh danganronpa music symphony hanging scholar tweaked regional bringing
Topic 10: symphony strike bolts brought military increasing lore dna chassis
Topic 11: admire comically hole wooden regional bringing starships basics scholar
Topic 12: hanging meadows minigames tires tired nigh succession idarb viable
Topic 13: nigh kids electricity holy hole basics regional bringing starships
Topic 14: basics bringing geography scholar viable hero tweaked regional starships

Neither the below average writers topics nor the random sample topics appears to have immediately identifiable groupings or categories. There are some specific game names and game terms in some of the topics (“idarb” was a recently released video game), but no game or genre of games seems to be appearing in either set.

Jargon Distance

We then tried to employ Jargon Distance calculations on the reviews based on score ranges to see if some writers use the same style or lexicon when talking about good or bad games. Jargon Distance was first proposed by Vilhena et al. to analyze communication and cultural holes between different scientific fields of study [1]. By computing Shannon and cross entropies of terms in a given set, efficiency of communication can be calculated as well as the jargon distance. Our implementation of this method borrows heavily from the method Vilhena et al. have outlined in their paper.

We first grouped all reviews based on their score by 10 point ranges (0-10, 10-20, 20-30 etc.) based on the hundred point scale. Of the ten possible score ranges, three of the ranges had no reviews and were removed from the analysis. We calculated the jargon distance between each group compared to every other group and averaged those values together to find the jargon distance between one range group compared to all the rest. Below are the average jargon distances for each range of scores.

Score Ranges	Average Culture Hole of Each Group with All Other Groups
group 1 ranges from scores in (10,20]	0.603877311999
group 2 ranges from scores in (40,50]	0.354558169854
group 3 ranges from scores in (50,60]	0.232642618479
group 4 ranges from scores in (60,70]	0.222251450889
group 5 ranges from scores in (70,80]	0.218852374037
group 6 ranges from scores in (80,90]	0.243247576346
group 7 ranges from scores in (90,100]	0.389665200183

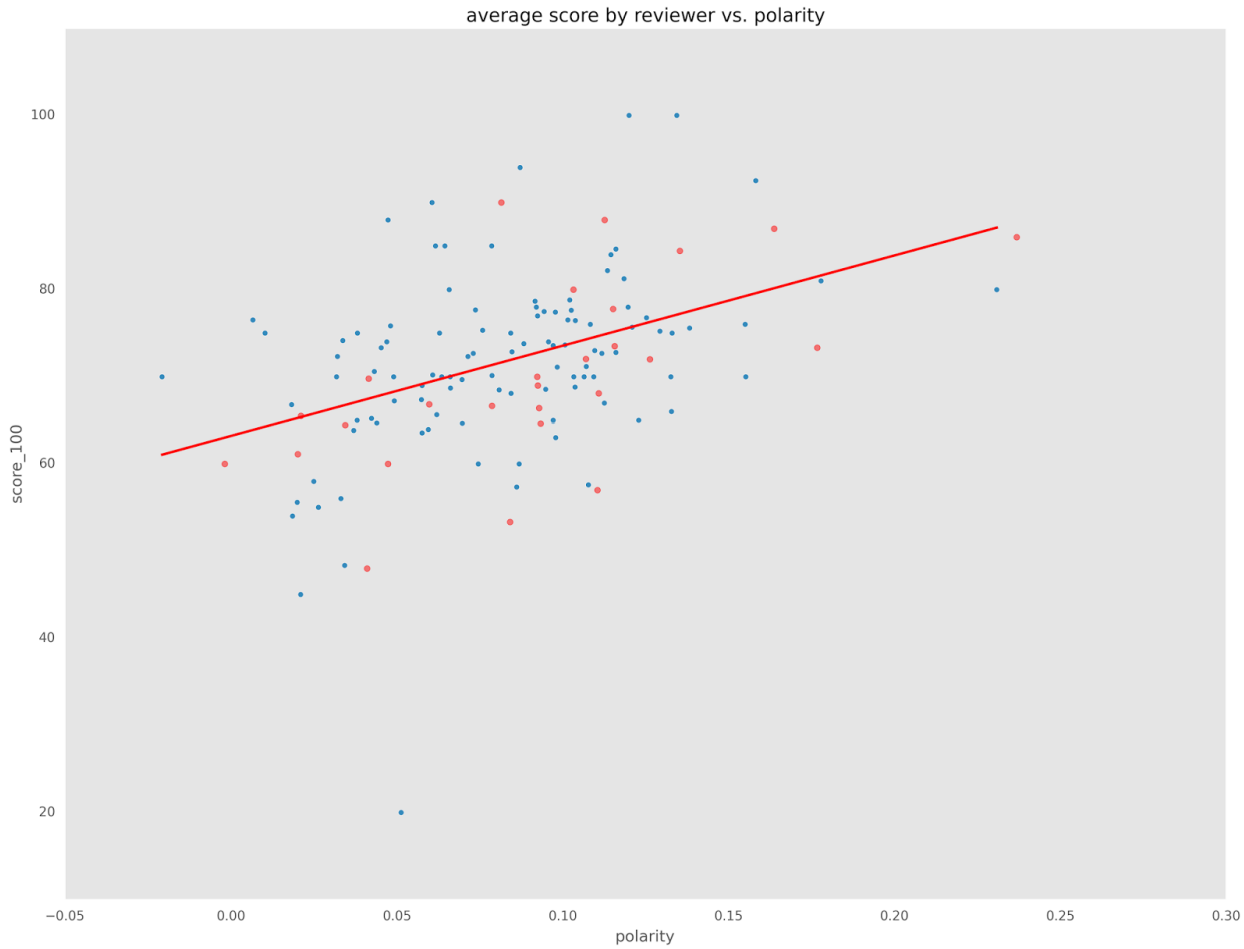
Table 1: Average jargon distance for each score range

Of note, groups 1, 2, and 7 have much higher values than the four other groups. However, group 1 has fewer reviews in its sample so it is likely that the sample size is creating a high number. Group 2 and 7 may have a higher score because the language used to describe a great game or a mediocre game is different than describing a game that ranges from acceptable to good. This is also reflected in the similarly in groups 3-6 with the values being so close.

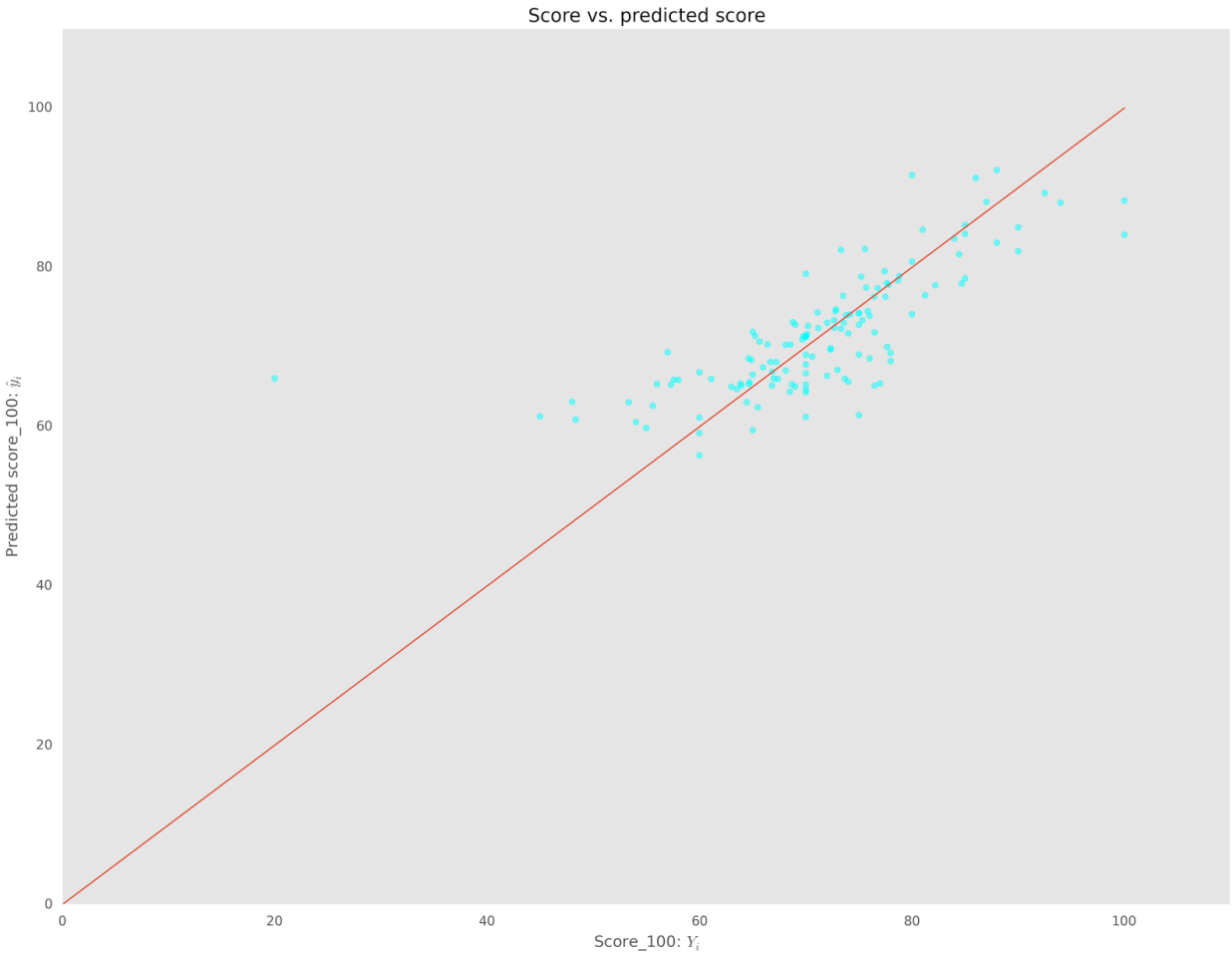
Seeing this closeness, we then attempted sentiment analysis to see if writer vocabulary changes for score.

Sentiment Analysis

We used the pattern analysis toolkit via the TextBlog package in python to calculate polarity and subjectivity scores on both processed and unprocessed reviews. After removing stopwords and punctuation, each review was treated with the pattern analysis algorithm to give a polarity score between -1 and 1. A 1 indicates that the paragraph is highly positive in sentiment. Each review also produced a subjectivity score between 0 and 1, which 1 being most subjective in language used. Using these values, we created a 20-80 testing and training set and ran an ordinary least squares linear regression on the review's printed score out of 100 against polarity score.



For the dataset, polarity stayed within the range of -0.05 and 0.25, meaning that each review was generally positive in sentiment. There appears to be is a slight positive correlation between sentiment polarity and the final review score, although with a data was likely overfit. R-squared over the training set was 0.154, and our p-value was 0. Adding additional features helps make our model more accurate. When accounting for the site the review was written on, the subjectivity, the review length, and the polarity, we can almost make predictions of what the reviewer's next potential score might be.



Discussion

While we had high hopes for our textual analysis to reveal patterns among writing styles and scores, our findings indicated that we only really scratched the surface of this expansive dataset. We used the entire corpus of text reviews mostly out of necessity as we lacked the additional metadata about each game specifically such as genre and theme. As such, our LDA method applied on such a large and specific set of text may have been responsible for producing insignificant results. Breaking up games into certain blocks and then running LDA to find common terms may be a better way to run topic modeling. However, our jargon distance and sentiment analysis results did have interesting results that could be used in a larger predictive model or recommender system.

To further this analysis, more analysis would have to be done on not only different genres of games but also on different gaming platforms as well. This could establish whether certain types of games review higher or lower compared to other genres and could determine if certain writer have higher average scores because they review certain types of games repeatedly. If a

recommendation system were to be created, some of the scoring logic from the analysis could be useful in normalizing scores from high and low genres.

Some of the interesting ideas we have for future work include training a nearest neighbor or regression model that taken into account the following variables: game genre (RPG, strategy, mobile, etc) and game platform (xbox, PlayStation, etc). These factors may contribute to the recommender system we had in mind when we began the research.

Conclusions

This project served as a starting point for video game review text analysis and found that there are unique findings that could be useful for writers to know as well as consumers who may rely on these reviews for purchasing decisions. By scraping text reviews from the internet and using text analysis packages like LDA, jargon distance, and sentiment analysis we were able to see some trends around writer score assignments and textual similarities between certain score ranges. While there is more research that can be done in this area, our early findings can provide a basis for further study in this new field of art and entertainment.

References

1. Vilhena, Daril A., Jacob G. Foster, Martin Rosvall, Jevin D. West, James Evans, and Carl T. Bergstrom. 2014. “Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication.” Sociological Science 1: 221-238. DOI: 10.15195/v1.a15