Predicting Video Game Sales Using Metacritic Reviews: Reviewers vs Users

anonymous
location
address 1
address 2

Abstract

write me!

AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions.

Introduction

Product reviews are often used by consumers as a purchase criteria. Particularly in the entertainment industry the opinions of an elite set of critics and journalists are often held as particularly important indicators of product quality. Professional film critics such as Roger Ebert are regarded as important guides to navigating the mass of films being released to recognize both hidden gems and over-hyped flops. Yet despite this widespread regard there is relatively little understanding of how indicative these critics' opinions are for market success of products. How effective are critic reviews for predicting future product sales? How do critic opinions compare to the opinions of regular consumers?

We address this question by comparing professional critics and amateur review website users in the task of predicting videogame sales. Are features of professional or amateur reviews more predictive of videogame sales? How does the power of grassroots word-of-mouth compare to the professional critical analysis? Do critics and users differ in their expressions of sentiment and does this effect their predictive power? How do professionals and amateurs describe products when reviewing? While previous research has often investigated the question of predicting sales from reviews, these efforts have focused on books (Gruhl et al. 2005) or films (Dellarocas, Zhang, and Awad 2007) (Yu et al. 2012) (Duan, Gu, and Whinston 2008) (Liu 2006), with little attention to the domain of videogames. Further, few of these analyses have explicitly compared professional and amateur reviewers. Our analysis addresses a question at the heart of many contemporary debates: how does the crowd compare to domain specialists and where and when are each effective for predicting future outcomes?

For our analysis we gathered a corpus of 197,383 total critic and user reviews from the review aggregation site Metacritic and sales data on 7467 games from the videogame sales tracking site VGChartz. Metacritic produces an aggregate "metascore" reflecting a proprietary weighted combination of professional reviewer scores. Anecdotally, many have claimed metascores can predict videogame success, at least when games cross a threshold of high enough score. First, we present an overview of our corpora that characterizes the distribution of professional and amateur review scores and sentiment expressions along with game sales. We use linear regression models to assess the predictive power of professional and amateur review scores for game total sales to date and sales over the first 10 weeks from being released. Our analysis includes using all review information and only a subset of reviews given prior to release of a game. Finally, we train classification models on review texts to find key words differentiating professional and amateur reviews showing different language and interests of the two groups. We find:

1. user are skewed toward more positive review scores than critics, but users provide the main source of low review scores
2. users and critics show very similar levels of subjectivity and review polarity
3. a combination of mean critic score, volume of critic reviews, and volume of user reviews accounts is effective for predicting game sales
4. critics and users systematically differ in their reviews, with experts focusing on market and sales factors and users focusing on social aspects of playing the game
5. metascores are less important predictors than the volume of critic or user scores in a model combining information from metascores, critic, and user scores

Related Work

The increasing prevalence and ease of access to social media data has led to a rise in research predicting real-world economic, political, and other events using (Asur and Huberman 2010). Predicting books sales (Gruhl et al. 2005), film box office sales (Dellarocas, Zhang, and Awad 2007)(Yu et al. 2012), and (to a lesser extent) videogame sales from online

reviews or blog chatter have all attracted attention (Ehrenfeld 2011)(Marcoux and Selouani 2009). In early work in this area, Gruhl et al. demonstrated blog chatter can predict spikes in book sales rank on Amazon (Gruhl et al. 2005). Predicting box office sales has found volume of mentions on blogs or Yahoo!'s movie prerelease discussion forums can predictive film box office sales (Liu 2006)(Duan, Gu, and Whinston 2008). Results conflict, however, on whether review score averages are (Dellarocas, Zhang, and Awad 2007) or are not (Liu 2006)(Duan, Gu, and Whinston 2008) predictive of sales, beyond the power of pure volume of reviews. Sentiment has also often proven useful in predicting sales. Yu et al. found sentiment expressed in blog posts can improve an autoregressive model through a carefully chosen lexicon of sentiment-laden words (Yu et al. 2012). Specialized classifiers that differentiate product description text from opinion related text were shown to improve sales prediction by Ghose and Ipeirotis (Ghose and Ipeirotis 2007). We address the question of review subjectivity and its relation to sales prediction and reviewer expertise, finding no strong predictive powers in the domain of videogame sales.

Few of these early works explicitly examine whether and how different types of reviewers produce varying review scores. Dellarocas et al. note a weak correlation between user and critic film reviews and find that removing critic review information (average score only) is slightly less detrimental to their model's performance than removing user review information (average score, number of reviews, entropy of review gender distribution, entropy of reviewer age distribution) (Dellarocas, Zhang, and Awad 2007). We provide a more detailed assessment of the differences of critic (professional) and user (amateur) reviewers through examining review volume, score, and sentiment rating in a linear model that allows direct comparison of predictive power. Further, our analysis of review text and score distributions provides more in-depth insight into how professional and amateur reviews differ.

zagal mention for text analysis

Our work broadens previous analyses of videogame sales that have been limited to smaller subsets of games or sources for review data. Ehrenfeld's thesis found volume of game mentions on an online game discussion forum (NeoGAF[1]) are predictive of game weekly sales in a support vector regression model. Marcoux and Selouani employed an autoregressive neural network model to predict games sales from review scores, volume, and related features using data from a single videogame news website (IGN [2]) after first performing nonlinear transformation on the model data. Compared to these approaches we examine a much larger set of games , compare professional and amateur reviews, and draw from a larger set of reviews. In contrast to previous non-linear models our results are directly interpretable in terms of relative feature meaning and impact according to the model.

ber
es

[1]http://www.neogaf.com/forum/
[2]http://www.ign.com/

Methodology

Corpus Collection

We used the python Scrapy package[3] to scrape all user and critic reviews for videogames from the Metacritic website[4]. For every game we collected the following information: console (the hardware the game software was made for), title, publisher (company responsible for distributing the game), developer (studio responsible for making the game), release date, current metascore, current average user score, genre (according to Metacritic), and ESRB rating (age-appropriateness rating). Note that some titles may appear on multiple consoles. We treat these as separate games as they reach potentially different audiences and may vary in their implementation. From every game we also collected all user and critic reviews, including their text, review score, time of review, and a flag indicating whether the review came from a user or critic. Metacritic converts critic review scores from many formats (such as letter grade, 0-10 range, 0-100 range) into a 0-100 score. User reviews are limited to a 0-10 score range. For comparison we divide all critic scores by 10 to have all reviews on a [0-10] scale. Metacritic only provides summary excerpts from critic reviews and we limited ourselves to this text to make review length more comparable between users and critics. Our final corpus consists of 197,383 reviews: 138,843 from critics and 58,540 from users. Of these, 45,679 critic and 58,531 user reviews had review date time stamps; this subset of reviews were used in our prediction tasks.

For sales data we scraped information from the VGChartz website[5]. VGChartz tracks game weekly, annual, and lifetime sales data from a variety of outlets and is primarily targeted toward sales of games in the United States. Their data is most accurate for console games (rather than computer or mobile phone), so we limit ourselves to examining the following consoles using only US sales data: Sony's PlayStation 3 (PS3), PlayStation Portable (PSP) and Vita; Nintendo's Wii, DS, and 3DS; and Microsoft's Xbox360. These consoles are considered to make up the current generation of videogame hardware and are the primary game console distribution platforms. For weekly data VGChartz typically only records the first 10 weeks of game sales, and so our time series analysis is limited to these weeks. We collected total lifetime sales-to-date for 7467 games and weekly sales for the first ten weeks of game sales for 4902 games.

Analytic Models

Our analysis involved two components: (1) predicting game sales (both lifetime and over the first 10 weeks of sales) and (2) identifying words distinguishing critic and user reviewers. Sales predictions investigated: (1) predicting total lifetime sales-to-date using the full review database, (2) predicting total lifetime sales-to-date using only reviews in the 10 weeks prior to the release of a game, and (3) predicting net first 10 week sales using the same pre-release review subset.

rename!

[3]http://scrapy.org/
[4]http://www.metacritic.com/
[5]http://www.vgchartz.com/

After matching game reviews to sales data using game titles and consoles we had lifetime sales data for 3649 games, and weekly sales for 600 games. We used a conservative approach of only keeping exact game title matches without manipulating titles (e.g. using lowercase versions or removing punctuation). In our case it is better to have a slightly smaller dataset than misattribute review scores to different sales. We predicted both full lifetime sales-to-date and summed weekly game sales data over the first 10 weeks of available information. Sales values were logged to ensure normality. Reviews and sales data were aligned based on weeks since release. Reviews were binned into weekly periods based on time since the release of the game. For example, all reviews within 7 days of the game's stated release date (for that console) were binned into the first week. We employed a simple linear regression model for better model result interpretability. Our model incorporated features for median review scores[6], number of reviews, median review text polarity, and median review text subjectivity.

Our analysis of review text examined the predictive power of review text words to distinguish reviewers as critics or users using penalized binomial regression with a lasso penalty to favor the use of a small set of terms. We prepared our review texts using several standard methods performed in text analysis using the R programming language packages tm[7] and topicmodels[8]. We removed whitespace, punctuation, numbers not part of words, and common English words (known as stopwords). All text was lowercased and stemmed to group together repeated use of similar words. We tokenized these documents into single words, requiring words be at least 3 letters long and appear in at least 10 documents. After constructing a full corpus of 86,376 terms we removed the most sparse terms to produce a set of 2,581 terms.

Binomial regression was performed using R's glmnet package[9] for generalized linear models. These models account for collinearity of terms (their frequent appearance together altering relative importance) and can control for sparsity (relative infrequency of terms). Controlling collinearity is important to prevent overweighting words that often appear together. Accounting for sparsity enables the model to exclude terms with little predictive power, yielding a smaller and more interpretable set of results. This was done using a lasso penalty that encourages the model to use the smallest set of terms possible

We employed python's pattern package[10] to analyze review text sentiment. For every review text we parse the text into sentences, compute per-word sentiment (both polarity and subjectivity) and compute the per-review mean, median, and maximum sentiment values. Subjectivity values measure to what extent a text conveys an opinion. Polarity values estimate that valence as positive or negative. Below we refer to these per-sentence aggregation values as the review "mean", "median" or "maximum" polarity or subjectivity.

[6]While we examined average review scores, these were found to always perform more poorly than median scores.

[7]http://cran.r-project.org/web/packages/tm/

[8]http://cran.r-project.org/web/packages/topicmodels/

[9]http://cran.r-project.org/web/packages/glmnet/

[10]http://www.clips.ua.ac.be/pages/pattern

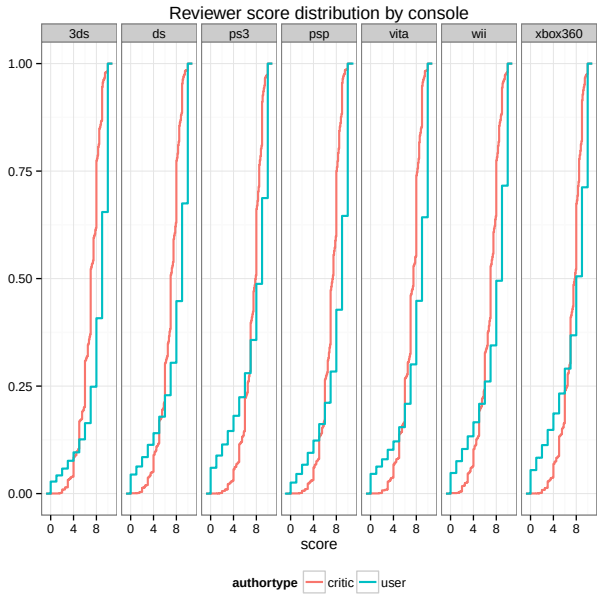


Figure 1: Cumulative review score distribution by console

Results

Corpus Overview

Before describing results on our prediction tasks it is important to understand the characteristics of the review and sales corpuses we collected. Metacritic users are not necessarily representative of the opinions of all those interested in videogames and it is not apparent *a priori* how they behave compared to the critics whose reviews are featured on the site. Users and critics show clear differences in review score attribution (Figure 1). Critics tend to provide reviews distributed more tightly around a mean 7.2 and median 7.5 score, while users show more variation with a mean 7.4 and median 9.0 score. A Wilcoxon test found these differences to be significant ($p < 0.001$). Also note that users favor providing higher scores than critics, but are the only ones likely to provide low review scores. Intuitively these results make sense. Critics are often under threat of blacklist for providing low review scores and have a reputation to preserve by not consistently giving high review scores. Game distributors are also unlikely to provide reviewers with free game copies for review if they anticipate low ratings, while critics are unlikely to review low profile and low quality titles. These factors combine to skew critics towards reviewing games generally favorably without providing overly positive reviews. Users, in contrast, are most likely to review when a game provided a great or terrible experience. As Metacritic is a major game review outlet reviewers are likely to provide high scores to games they enjoyed, while attacking games they found poor quality or a waste of money.

	critics		users	
	mean	median	mean	median
score	7.2	7.5	7.4	9.0
mean polarity	0.14	0.12	0.15	0.13
median polarity	0.14	0.10	0.12	0.00
max polarity	0.25	0.22	0.51	0.50
mean subjectivity	0.50	0.50	0.48	0.48
median subjectivity	0.50	0.50	0.47	0.50
max subjectivity	0.65	0.69	0.86	1.00

Table 1: Comparison of critic and user review text sentiment.

Reviewers and critics also show different levels of subjectivity and polarity in reviews (Figures 2, 2 and 3). For our analysis we consider median polarity and subjectivity values over all review scores; all results reported were significant according to a Wilcoxon test at $p < 0.001$, although these differences may not appear in Table when rounding to two decimal places. Both critics and user review distribution are slightly skewed by the presence of larger numbers of high polarity and subjectivity reviews (compare means and medians in Table). Critics tend to give lower scores and show less skew in review scoring—7.2 mean and 7.5 median for critics, 7.4 and 9.0 for users. Critics thus tend to be more consistent in scoring and more conservative about providing overly positive reviews. Users are more likely to employ at least one sentence that is strongly subjective and of more positive valence than critics. Both users and critic reviews are skewed towards moderately positive sentiments, although most cluster around neutral tones (Figure 2).

While differences in polarity and subjectivity were significant, their magnitude was small. Contrary to our expectations, critics are not particularly objective when compared to users, except in terms of the most polarizing sentences used in their reviews. One limitation of this interpretation, however, is that critic texts were limited to summaries and thus may not reflect the intended valence of the review, but a revised summary meant to convey factual information for the purposes of featuring on Metacritic. These results may also reflect limitations of our sentiment analysis itself and merit further investigation using full review texts from the original review sites.

Sales Prediction

We used a linear model to predict mean lifetime sales for each game. While linear models may have limited predictive power for nonlinear relationships they afford direct interpretation useful to analysts employing the results. We examined a set of models for predicting game sales-to-date and net first ten week sales considering a set of factors: review author (critic or user), review scores, number of reviews, and review text features (length, polarity, subjectivity). We examined three models: predicting total sales from all reviews, predicting total sales from reviews in the 10 weeks prior to the game's release, and predicting net first 10 week sales from the same pre-release review subset. Review scores and volume show clear relationships to the lifetime sales of a game (Figures 4 and 5).

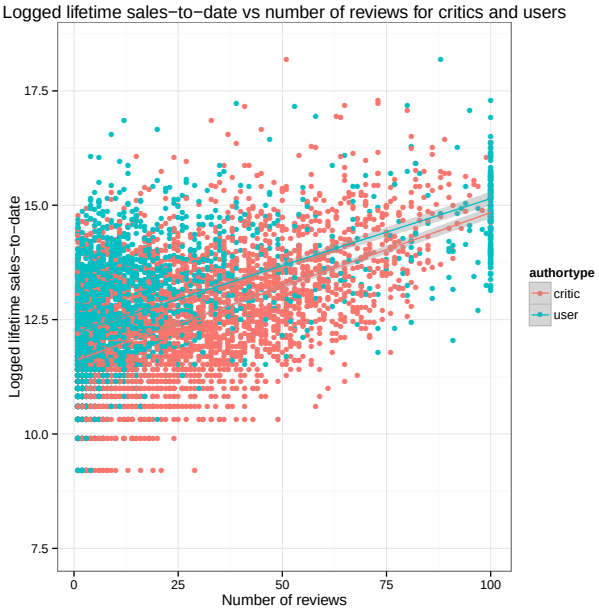


Figure 4: Lifetime sales-to-date vs number of reviews by reviewer type

Figure 5: Lifetime sales-to-date vs median review score by reviewer type

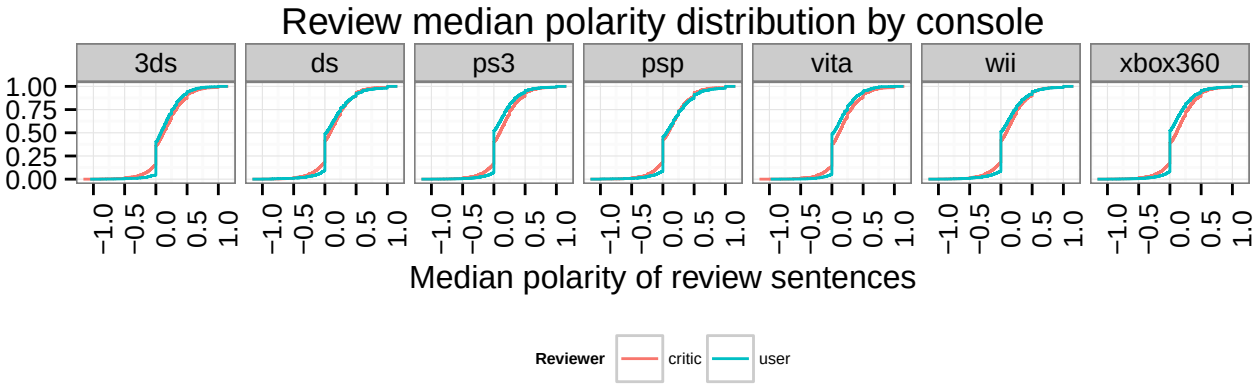


Figure 2: Cumulative review polarity distribution by console comparing reviewer type

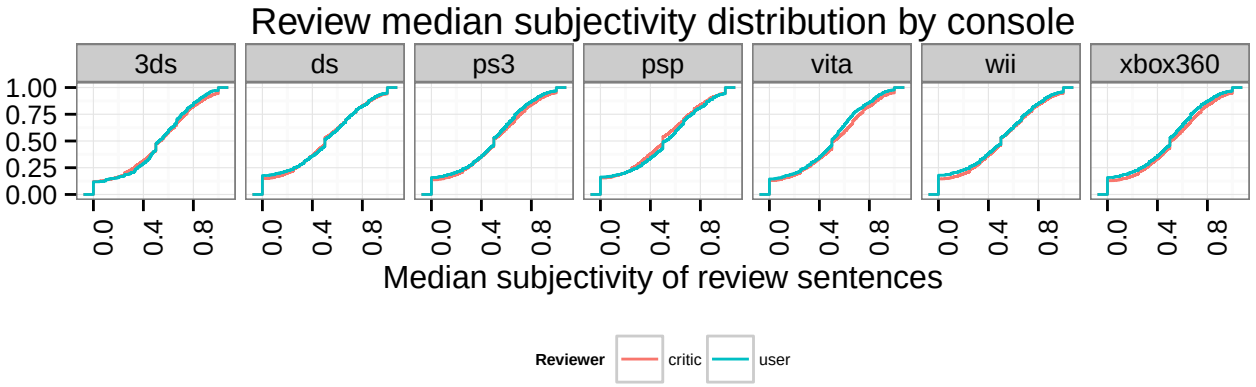


Figure 3: Cumulative review subjectivity distribution by console comparing reviewer type

Anecdotally many have claimed that metascore over 90 (9 on our rescaled range) leads to a substantial boost to game sales. Our data supports this notion when looking only at aggregated critic scores, although the range of relevance is broader (Figure 6).[11] User scores do not show a clear relationship to sales: increasing median user review score does not appear to relate to any substantial gain in scores. By contrast, critic review scores show a clear increasing relationship, starting around scores above 60 but becoming most prominent for scores above 80. While the [95-100] category has only 25 critic reviews and thus is of dubious predictive value, the [85-90] and [90-95] categories consist of 371 and 56 reviews and thus are of greater predictive merit.

Total sales vs all reviews Our regression models show moderately strong predictive power (R^2 between 0.275 and 0.390)for logged sales; Tables and summarize our results. All regressors were converted to standardized scores by subtracting their means and dividing by their standard deviations. Regression results report only significant ($p < 0.001$) standardized β coefficients.

For each model we computed all possible linear models from combinations of review score and text features and scored them according to the Bayesian Information Criterion (BIC) using R's leaps package.[12] Using the BIC allows our model selection process to penalize overly complex models to reduce the likelihood of overfitting. Relative importance of model parameters was measured using the lmg metric in R's relaimpo package.[13] The lmg metric calculates the average R^2 contribution averaged over orderings among regressor variables—we use relative values among regressors to compare their importance.

For total sales we examined a model treating critic and

[11] We found similar trends when examining Metascores, although the dataset is limited to fewer games.

[12] http://cran.r-project.org/web/packages/leaps/

[13] http://cran.r-project.org/web/packages/relaimpo/

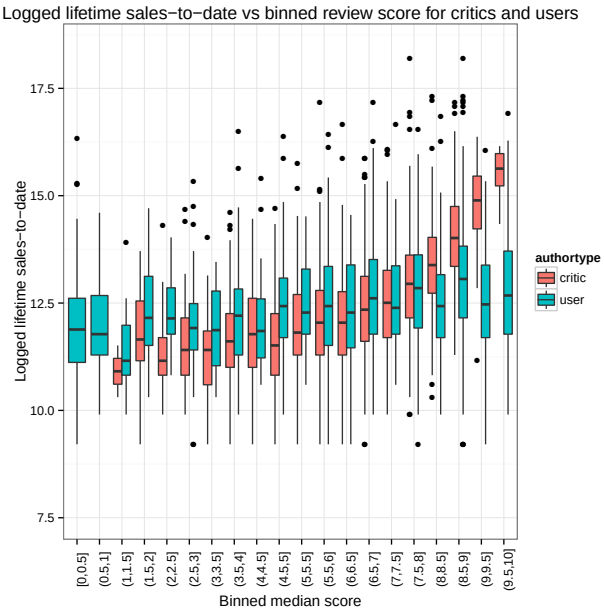


Figure 6: lifetime sales vs binned review scores

	total vs all reviews ($R^2 = 0.390$)	
variable	β	**relative importance**
mean critic score	3.084e-01	30.85%
mean user score	-6.273e-02	2.53%
number critic reviews	1.729e-01	27.49%
number user reviews	3.165e-01	37.43%
mean user review length	-4.592e-02	0.49%
median(max review subjectivity) critics	-6.431e-02	0.86%
median(max review subjectivity) users	-4.462e-02	0.36%

Table 2: Regression models for lifetime sales-to-date using all available reviews. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values are significant at $p < 0.001$.

user reviews as separate sources of regressors, looking only at games with data from both groups and using all reviews in our dataset (2909 games). The BIC identified the following variables (relative importances summing to 100% across regressors are reported in parenthesis): mean critic review score (30.85%), mean user score (2.53%), number of critic reviews (27.49%), number of user reviews (37.43%), mean user review length in characters (4.92%), median of per-review maximum subjectivity from critic reviews (0.86%), mean of the same from user reviews (0.36%). Overall the model's adjusted R^2 value of 0.3904 suggests a strong correlation between these factors and lifetime sales.

As expected from the previous distributions, review subjectivity is a weak factor for predicting total sales. Sheer volume of user and critic reviews account for the largest part of our model's power. Intuitively this makes sense: volume of critic and user reviews both reflect relative attention to a game and also likely reflect the relative publishing and marketing resources behind the game's developer. More user reviews reflect games more people have played (and thus likely already purchased). Further, games that drive users to respond on online forums reflect greater social popularity. Greater volume of critical reception similarly reflects publisher budget and expectations. Review copies are typically distributed to critics when publishers expect games to be reviewed well, bolstering their review numbers. Further, only more well-financed and established publishers are able to push their games to be reviewed by venues that metacritic would index and report from.

While critic scores and number of reviews have roughly equal relative importance, user scores are relatively unimportant. Given the distributions seen above this makes sense: games that have many user reviews are rare, but enjoy strong sales on average. For a game to have many user reviews requires many users to have played the game and been motivated to review the game. From the previous distributions of review scores we would expect these reviewers to (mostly) give positive scores, thus the fact that they are writing a score at all is indicative of what that score will be. Playing the game typically means a user has purchased the game, thus further entangling with sales. While direct causality cannot be read off from these results it is clear that the amount of attention users or critics devote to a game is indicative of its level of sales success. Future work should explore more sophisticated features for predicting sales such as reviews in the weeks prior to launch, number of reviews in the first weeks of sales (not just overall), or volume of previews. Twitter, Facebook, and Youtube and other media consumption information (not necessarily reviews) may be important gauges of broad consumer interest that complements the more devoted group of Metacritic user reviewers.

Total and net 10 week sales vs pre-release reviews For the purpose of forecasting future sales we examined a model that aggregated reviews only in the 10 weeks prior to the release of a game and used this data set to predict lifetime sales-to-date. Due to sparsity of games (600 total) that possessed both critic and user reviews in this period we altered the form of the model to use the same predictive variables as above but with a dummy variable indicating review type and adding terms to assess interaction effects between this variable and the other regressors. With an adjusted R^2 of 0.275 the model is less powerful than before. Relative variable importance mirrors the prior results: number of reviews (42.78%), median review score (31.59%), reviewer type flag (20.63%), interaction effect between reviewer type and number of reviews (1.18%), interaction effect between reviewer type and median score (3.83%); interaction effects between mean of per-review mean polarity and mean of per-review mean subjectivity were both unimport (0.00%). Directionally these results have the same message: volume of reviews is most important, followed by receiving favorable review scores, with all other factors playing a relatively minor role. Knowing users are providing review scores lowers the estimated effect on sales while knowing more users are writing reviews raises the estimate effect on sales. We created an analogous model predicting net sales over the first 10 weeks,

	total vs all reviews ($R^2 = 0.303$)	
variable	β	**relative importance**
metascore	2.908e-01	26.84%
median critic score	-1.338e-01	3.54%
number critic reviews	3.151e-01	41.62%
number user reviews	1.446e-01	27.27%
median(median review subjectivity) for users	-5.038e-02	0.73%

Table 4: Regression model for lifetime sales-to-date comparing metascores, users, and critics. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values are significant at $p < 0.001$ except median(median review subjectivity) which is marginally significant at $p < 0.1$.

achieving an R^2 of 0.275. Results were qualitatively the same as the total sale model, refer to Table for details. This is not surprising as most game sales occur during the initial launch and marketing period, with relatively few games having more long-term strong sales or shifts in sales.

Comparing Metascores, critics, and users To better understand the powers of different scoring systems we compared the predictive power of metascores, combined critic review scores, and combined user scores for predicting lifetime sales-to-date. From the total dataset, 840 games have data on lifetime sales-to-date, Metacritic metascores and reviews from both critics and reviews. We predict lifetime sales as metascores are running values that are updated without a historical trace—we thus cannot know historical values of metascores prior to a game's launch. Critic and user reviews were aggregated across the full set of data, rather than limiting to the pre-launch subset to allow for a fair comparison.

According to the BIC criteria the best linear model ($R^2 = 0.303$) uses: metascores (26.84%), median critic scores (3.54%), number of critic reviews (41.62%), number of user reviews (27.27%), and the median of the median review subjectivity from user reviews (0.73%) (Table). Median user review subjectivity was only marginally significant ($p < 0.1$) while all other coefficients were significant ($p < 0.001$). As in the previous models, volume of user and critic reviews captures a large portion of the model's predictive power, using metascores for most score information. These results suggest metascores are more predictive than user or critic scores combined by simple weighting schemes (mean or median). However, without historical information on metascores it is impossible to discover whether metascores are a leading indicator of sales, lagging result, or simply strong correlate due to the external market factors mentioned above. The comparable overall model power between using only user and critic data for lifetime sales and adding metascores suggests the metascore weighting scheme may not be particularly powerful compared to simpler methods.

ranking reviewers no users w/more than 10 reviews + strong sales correlation. some critics negative correlation. subset of top gets prediction to ≈ 0.308.

review "importance" via score correlation; not needed for main story

Review Text Classification

Our previous analyses show reviews are predictive of sales to some extent. But what aspects of reviews distinguish them, particularly in terms of the words used? We explored this question by using review text in a bag of words model to classify reviewers as critics or users using a penalized binomial regression model. Understanding these text-level differences can enable better prediction of what characteristics distinguish critic reviews in content from user reviews and how these text-level differences express general trends in how critics and users review.

update me

The model achieved 90.71% accuracy over the set of 197,383 reviews we had data for. 96.38% of critics were accurately labeled, and 77.28% of users were correctly labeled, reflecting the imbalanced proportions of these categories in our data set—70.34% of reviews come from critics, 29.66% from users. The penalized regression model employed 2,570 terms, assigning 146 a weight of zero, 1235 positive weights, and 1190 negative weights. Thus, 146 terms had no power to predict whether a reviewer was a user or critic, while 1235 terms were positively associated with a reviewer being a user and 1190 terms were associated with a reviewer being a critic. Below we examine several of the terms from these groups to understand how users and critics differentiate themselves in text descriptions. We report term coefficients in parenthesis next to terms to help understand these analysis. As words were stemmed we have included completions of the stems in parentheses to help interpretation.

update me

Critics were most recognizable by their use of month names and holidays in reviews, followed by references to the games market and gameplay feature descriptors. Compared to users critics are more likely to mention words linking a game to marketing and game industry competition, using terms such as "tie(-)in" (-1.53) and "competitor" (-1.44). References to game features included descriptions of gameplay perspective— "firstperson" (-0.86) and "thirdperson" (-1.01)—as well as gameplay style–"action" (0.36), "roleplay" (-1.34), "actionpack(ed)" (-1.21), and "brawler" (-0.78). Critic reviews thus seem to be characterized by an effort to characterize a game within the broader industry sales cycle and specific genres of interest. As critic reviews are often a form of marketing for companies this makes some intuitive sense—critics are acting to define the intended audience of a game through describing how it relates to a broader marketing strategy and genre.

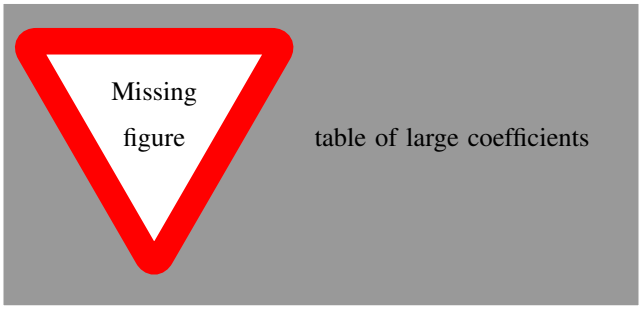
Users were recognizable by slang terms and references to gaming culture and reviews. Slang terms included phrases such as "wtf" (2.32), "imo" (4.59), "lol" (2.77), and "meh" (1.20). References to gaming culture often mentioned the culture of reviewing and purchasing: "bought" (1.92), "review" (1.74), "opinion" (1.46), "bias" (2.45),

	10 week vs pre-release ($R^2 = 0.306$)		**total vs pre-release** ($R^2 = 0.275$)	
variable	β	relative importance	β	relative importance
median review score	0.6993	32.46%	0.5998	31.59%
number of reviews	0.2438	56.36%	0.2249	42.78%
reviewer type (as user)	0.4637	4.55%	0.6248	20.63%
median score X reviewer type (as user)	-0.7026	4.93%	-0.5840	3.83%
number of reviews X reviewer type (as user)	0.6918	1.71%	0.5368	1.18%
reviewer type (as user) X mean(mean review polarity)	0.2306	0.00%	0.2005	0.00%
reviewer type (as user) X mean(mean review subjectivity)	-0.1898	0.00%	-0.1604	0.00%

Table 3: Regression models for lifetime sales-to-date and first 10 weeks net sales using reviewers in 10 weeks prior to game release. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values are significant at $p < 0.001$.

"overr(ated)" (4.42), "underr(ated)" (2.84), "metacrit(ic)" (3.03), "gamestop" (3.31), "ign" (2.01). Gamestop is a major game distribution store chain and IGN is a major game-related news and gameplay website. Users were also characterized by referring to specific features of playing games: "troll" (2.83), "fanboy" (1.30), "hater" (0.94), "wife" (2.74), "son" (2.06), "server" (1.40), "lagg(y)" (1.27), and "beta" (1.65). Servers, lag, and beta all refer to common aspects of networked online play; fanboys are fanatic fans of a game franchise or console. Mentioning both live gameplay features and family members reflects users' stronger emphasis on the experience of playing with others, rather than assessing a potential product for (implicitly individual) consumption.

Terms removed from the model varied widely, covering topics that in some ways bridge the two reviewer groups. Several terms referenced games as franchises—"reboot" and "sequel"—while others noted target audiences—"child" and "adult". "Broken" a term referring to dysfunctional gameplay seems to be shared across users and critics. Overall, however, there do not seem to be strong patterns among words not distinguishing the groups, suggesting these terms are simply a general vocabulary with relatively little significance for understanding game reviewing behavior.



Combined, these results paint a picture of critics taking a more professional role of identifying purchase products and describing their features, while users relate their games to broader consumption practices and gaming culture. Users are free to reference particular game distributors or review biases in ways critics cannot. Compared to critics, users typically relate games back to playing behavior, experiences, and social interactions. In part these results help understand the power of critic review scores to predict sales better than user review scores. Critics describe games in a way to guide purchasing decisions, while users are more likely to reflect on a game in their play practices and purchasing experiences with relatively little information to help understand the product features themselves.

Discussion

rewrite

Our analyses of game sales and reviews uncovered several primary differences between users and critics that highlight their differential roles in promoting and consuming games. Critics employ a narrow range of review scores, describe games in terms of product purchase decisions, and are predictive of lifetime sales both in terms of scores assigned and volume of reviews. Combined, these results paint a portrait of critics as dispassionate professional sources of information and indicators of overall investment in game production values and marketing push. In contrast, users are more likely to give low review scores, tend to skew towards high review scores, make reference to gaming culture and experiences around playing a game, and are only predictive of sales through the volume of reviews provided. Together these results portray user reviewers as consumers evaluating the quality of a game experience and acting as indicators of overall market uptake and attention to the product.

Comparing users and critics we see critics as providing "objective" facts describing games, with scores guiding initial market interest. User reviews appear to gauge the broader uptake of games by a community of players, with scores indicating the relative success or failure of a game to meet expectations, although too late to have a strong negative impact on sales of the given game (although perhaps not the franchise). Overall, users and critics provide alternative perspectives on games through their reviews, with both correlating with future sales success.

Limitations

not sure about structure

Our work has several important limitations both in terms of the models employed and generalization of our results. We intentionally employed linear predictive models for lifetime sales and running weekly sales. Linear models allow for easy and direct interpretation but lose the nuance in these al-

most certainly non-linear relationships. These complications highlight a limitation in interpreting our results: while clear relationships exist in our data they do not necessarily indicate causality and likely reflect a host of influencing factors. Game unit pricing and sales, marketing, developer and publisher renown and external factors around sales regulation and economic circumstances all impinge on game sales. Further, we are limited to third-party aggregation data and thus have only limited accuracy in sales reporting. That our models have the strong correlations they do is impressive in light on these factors.

Our combination of game data into relative time since launch prevents us from accounting for seasonal differences in sales - such as holiday sales jumps - or interactions between sales of games released at similar times. Sales can easily be hurt by prominent competitors or bolstered through cross-marketing with other games, factors we currently ignore.

The data we used potentially limits generalization based on our model. Our corpus is derived from a single website and critic reviews were limited to summary text. It is unclear whether the user/consumer reviewing patterns of users on other websites would show similar relationships as Metacritic is considered a prominent source for game review information. Perhaps the number of user reviews only matters from major review outlets, while other venues have less influence on sales. Limiting critic text to summaries may have hurt the power of our models to detect any language critics employ in a full review text. Yet this is typically what users would find when seeking critic reviews. By employing summaries we maintained the interface of Metacritic, but this comes at the cost of missing the more general patterns of how critics review when presenting materials on their own venue.

Our text-based models are further hamstrung by using solely unigrams in a bag-of-words model. The prominence of terms like "hater" as a positive association with review score highlights the role of linguistic qualifications and negations in reviews. The terms we provide are an initial foray into the space of review descriptions, but further work is needed to untangle the kinds of relationships being described. Even capturing simple adjective-noun relations would add depth and potential insight into what aspects of games reviewers describe and how they describe them .

tion Za-

Applications

The most obvious application for our work is predicting game sales from Metacritic reviews. Publishing firms can use this as a guide to forecast sales based on ongoing success to vary the resources devoted to supporting a game (particularly those with online or ongoing components). Marketers might attempt to drive user review volume and word-of-mouth in an effort to improve early and lifetime sales. Although we have not demonstrated a causal connection, this strategy presents itself as one worth exploring.

Review sites such as Metacritic develop ad hoc schemes for converting among multiple review systems. Employing predictive models based on score or features could potentially provide a unifying metric for game "quality" that puts all scores on a unified scale. For game marketers and producers this can quantify the value of reviews they get when selecting potential reviewers. For reviewers, this can enable them to focus reviews on their desired features without having to conform to accepted scoring systems. Losing familiar user guides to scoring may hurt interpretability for users, but may also allow both users and critics to focus more on aspects of game features or experience without filtering these results through an arbitrary number.

Future Work

Several avenues are available for extending this work toward better understanding the user-critic distinction and improving the predictive models employed. Beyond improving the sophistication of the natural language processing techniques used to examine review texts we might compare users and critics from other websites or sources. Do users of different websites provide different perspectives on games? Metacritic is known as a major review center and thus likely reflects the opinions of more devoted game players. Other review sources such as Amazon may have a wider audience that is indicative of general opinions or perspectives on games. Combining these sources may uncover how users of these sites differ and have varying predictive powers. For example, more "casual" games' sales may be better predicted by Amazon reviews while Metacritic reviews may better capture sales of more niche games.

Metacritic claims the metascore is a valuable summary of the quality of a game. Can we construct an analogous metric for weighting critic scores based on predictive accuracy for sales data? Can we employ review and sales data to assign relative value to critics based on their use for predicting future sales? We can even imagine constructing a "meta-metacritic" that combines the assessments of multiple review websites into a single metric predictive of game sales. Such a rating might convey how useful information from different websites is for assessing a game and predicting future success.

Given the predictive power of the sheer number of critic or user reviews prior to game release we might consider alternative ways to gauge this form of grassroots interest. Amazon or other major online retail sources' reviews may reflect more general market trends and levels of interest. As most reviews happen within the first two to three weeks of a game's release we could potentially gather further information on the "pulse" of interest using real-time information from Twitter, Facebook, and similar social venues. Do these sources provide complementary information? Can they enable better real-time prediction? Trends in Google search behavior (available from Google trends) or advertising Youtube videos may also provide useful supplemental predictive power (Asur and Huberman 2010).

Our analyses were limited to the first 10 weeks of game sales and excluded personal computer and mobile distribution platforms (such as Apple's app store or Google's Android marketplace). These venues hold promise for additional aspects of user reviews tied to the long-term success of a game. Console games traditionally have the bulk of sales early in release followed by a long tail of reduced

sales. Yet some titles are "evergreen" and continue to have comparatively strong sales over months or years of their life. User reviews in particular may be predictive of these titles as users often provide reviews much longer after a game's release compared to critic reviews that tend to cluster around release. Mobile and personal computer markets are also widely considered to be domains where this evergreen phenomenon is more common, making them valuable sources of information to consider. Expanding our analyses to these more longitudinal domains has great potential for improving our sales prediction methods and potentially identifying game traits users describe when choosing to play a game that is older. Developing more sophisticated models that account for the shape of the sales distribution over the first few weeks of sales—such as diffusion models (Dellarocas, Zhang, and Awad 2007)—may also hold potential for more effective and still interpretable results.

Expanding on the question of the power of the "wisdom of the crowd" we might compare alternative sources of aggregated knowledge such as predictive markets. The simsExchange[14] prediction market is one example used for predicting game sales. Comparing these sources against other critic opinions—such as professional industry analysts or company quarterly forecasts—may yield additional insights into how these groups differ and what factors correlate with sales.

Conclusions

We presented an analysis of user and critic reviews and examined their power to predict game sales. Pre-release reviews are moderately strongly correlated with early and long-term sales. Volume of response from both users and critics is most important, followed by critic (but not user) scores. Our work provides first steps towards understanding how

Acknowledgments

To be filled with non-anonymous information.

References

Asur, S., and Huberman, B. 2010. Predicting the future with social media. *arXiv preprint arXiv:1003.5699*.

Dellarocas, C.; Zhang, X.; and Awad, N. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing* 21(4):23–45.

Duan, W.; Gu, B.; and Whinston, A. 2008. Do online reviews matter?an empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Ehrenfeld, S. 2011. *Predicting video game sales using an analysis of internet message board discussions*. Ph.D. Dissertation, San Diego State University.

Ghose, A., and Ipeirotis, P. 2007. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, 303–310. ACM.

[14]http://www.simexchange.com/

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; and Tomkins, A. 2005. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 78–87. ACM.

Liu, Y. 2006. Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing* 70(3):74–89.

Marcoux, J., and Selouani, S. 2009. A hybrid subspace-connectionist data mining approach for sales forecasting in the video game industry. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 5, 666–670. IEEE.

Yu, X.; Liu, Y.; Huang, X.; and An, A. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering* 24(4):720–734.