Predicting Video Game Sales Using Metacritic Reviews: How do professionals and amateurs compare?

Abstract

Both professional critics and amateur users provide reviews on a multitude of products, potentially providing valuable information about future product success. Previous work has typically analyzed user-provided reviews from large sales websites or specialized interest group forums. We instead address how professional and amateur reviews differ and how effective these review sources are for predicting sales. Using the relatively under-studied domain of videogames, we examine a corpus of review texts from the popular review site Metacritic and correlate features of these reviews to future videogame sales. Compared to professionals, amateurs skew towards providing higher scores, are the primary source of low scores, but show nearly the same distribution of expressions of sentiment. A combination of professional scores, volume of professional reviews, and volume of amateur reviews in the 10 weeks prior to a game's release provides a moderately powerful predictive model ($R^2 = 0.472$, $MSE = 0.537$) for total game sales. Examining review texts, we find professionals favor phrases describing the videogame market while amateurs emphasize social aspects of playing games and the broader gamer culture. In sum, our work explores the differences between professional and amateur reviewers both in terms of their value as signals for product success and as representatives of different interest groups within the broad videogame product ecology. Our work supports pre-launch sales prediction for games based solely on existing review information already being aggregated in public review websites, leveraging the different contributions of professional and amateur reviews.

Introduction

> After paying $60 on this game I went home and put it in my PS3. After a couple hours of gameplay I realized I had been ripped off! I looked in my DVD rack and grabbed a [franchise] game I bought 2 years ago, it looked the same but had a 2 on it instead of a 3. I then realized it is the same exact game, with a few minor tweaks. Biggest waste of money ever. Yet, there are still trolls on the internet like [magazine]'s own [professional critic] getting paid big bucks to give it high review scores.

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 science on how indicative these professionals' opinions are for market success. How effective are professional reviews for predicting future product sales? How do professional 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 professional critical analysis? Do professionals and amateurs 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 (see (Gilbert and Karahalios 2010) for a study of differences among 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 reviews (138,843 from professionals and 58,540 from amateurs) from the review aggregation site Metacritic along with sales data on 7,467 games from the videogame sales tracking site VGChartz. Our data covers reviews from these sites through December 7, 2012. Metacritic collects reviews from print and online videogame review publications and produces an aggregate "metascore" reflecting a proprietary weighted combination of professional reviewer scores. Amateurs are able to provide their own reviews for the same games. 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 net sales over the first 10 weeks after release using reviews from the 10 weeks prior to release. We further compare the predictive power of specialized expert opinion aggregation systems ("metascores") against simple aggregations of professional and amateur reviews. 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. Amateurs are skewed toward higher review scores than professionals, but amateurs provide the main source of low review scores. That is, they show more variation.
2. Amateurs and professionals show very similar levels of average review subjectivity and polarity. However, amateurs tend toward more extreme and more varied expressions of sentiment.
3. A combination of mean professional score, volume of professional reviews, and volume of amateur reviews is effective for predicting game sales.
4. Metascores are less important predictors than the volume of professional or amateur scores in a model combining information from metascores, professional, and amateur scores. Oddly, metascores correlate very strongly with aggregate amateur scores, but only moderately correlate with professional scores.
5. Professionals and amateurs systematically differ in their reviews, with experts focusing on market and sales factors and amateurs focusing on social aspects of playing the game.

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 amateur and professional film reviews and find that removing professional review information (average score only) is slightly less detrimental to their model's performance than removing amateur review information (average score, number of reviews, entropy of review gender distribution, entropy of reviewer age distribution) (Dellarocas, Zhang, and Awad 2007). Early studies of review texts have characterized amateur reviewers in terms of their motivations for reviewing (Gilbert and Karahalios 2010) and characteristic language used for describing videogame products in particular (Zagal, Tomuro, and Shepitsen 2011). We complement this work through an assessment of the differences between professional and 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 in-depth insight into how professional and amateur reviews differ. Rather than characterize amateur reviewer traits in isolation we seek to understand how they compare against professionals.

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 the NeoGAF online game discussion forum is predictive of game weekly sales in a support vector regression model (Ehrenfeld 2011). Marcoux and Selouani employed an autoregressive neural network model to predict games sales from review scores, volume, and related features using data from the IGN videogame news and reviews website after first performing nonlinear transformations of the model data (Marcoux and Selouani 2009). Compared to these approaches we examine a much larger set of games, [1] compare professional and amateur reviews, and draw from a longer-term and larger set of reviews that aggregate over websites. Our results are directly interpretable in terms of relative feature impact in linear models.

[1] Ehrenfeld does not report the number of games used in his analysis, but limits to data over the course of 42 weeks of releases, which is a subset of our total corpus spanning 425 weeks. Marcoux and Selouani examine 74 games, while we examine 3000 or 600, depending on the model.

Methodology

Corpus Collection

Metacritic reviews We crawled all amateur and professional reviews for videogames from the Metacritic website[2] through December 7, 2012. 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 amateur 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 amateur and professional reviews, including their text, review score, time of review, and a flag indicating whether the review came from a amateur or professional. Metacritic converts professional review scores from many formats (such as letter grade, 0-10 range, 0-100 range) into a 0-100 score. Amateur reviews are limited to a 0-10 score range. For comparison we divide all professional scores by 10 to have all reviews on a [0-10] scale. Metacritic only provides summary excerpts from professional reviews and we limited ourselves to this text to make review length more comparable between amateurs and professionals. Our final corpus consists of 197,383 reviews: 138,843 from professionals and 58,540 from amateurs. Of these, 4,331 professional and 729 amateur reviews cover 958 games were in the first 10 weeks prior to the release of the game—these were used in our future sales prediction tasks.

VGChartz sales data For sales data we scraped information from the VGChartz website[3]. 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 7,467 games and weekly sales for the first ten weeks of game sales for 4,902 games.

Methods

Sales prediction regression Our analysis involved two components: (1) predicting game sales (both lifetime and over the first 10 weeks of sales) and (2) identifying words distinguishing professional and amateur reviewers. Sales predictions investigated:

1. predicting total lifetime sales using reviews in the 10 weeks prior to the release of a game (total vs pre-10)
2. predicting net first 10 week sales using the same pre-release review subset (10 week vs pre-10)
3. predicting lifetime sales comparing metascore, professional, and amateur review scores using all available reviews (total vs meta)

Predictive accuracy was assessed through mean squared error (MSE) computed with leave-one-out cross-validation through bootstrapped resampling with R's boot package[4]; Table 1 summarizes the model data sets and performance.

After matching game reviews to sales data using game titles and consoles we had lifetime sales data for 6,809 games, and weekly sales for 600 games. Of the 6,809 games with lifetime sales 2,902 had both professional and amateur reviews, while 839 additionally had metascores. Metascores are only assigned when at least four professional critics approved by Metacritic have reviewed the game.

We took a conservative approach of only keeping exact game title matches between Metacritic and VGCharz data 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 games. All games were matched by a concatenation of title and game console as games on different consoles vary in implementation, release date, and audience demographics.

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. Sales values were logged to get closer to normality. From reviews we constructed features for mean and median review scores, number of reviews, and review length. All features (including sentiment features below) were centered and scaled by subtracting their mean and dividing by their standard deviation.

Review sentiment extraction We employed python's Pattern package[5] 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, minimum, maximum, and variance of 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", "minimum", "maximum", or "variance" of polarity or subjectivity. "Overall" subjectivity and polarity refer to values scoring the review text as a whole, rather than computing per-sentence values and aggregating them. Aggregated these per-sentence features to compute per-review mean and median values of the above features. These multiple metrics allowed us to consider expressions of average review sentiment (median, overall), skew in re-

[2] http://www.metacritic.com/

[3] http://www.vgchartz.com/

[4] http://cran.r-project.org/web/packages/boot/

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

predicted variaable	**review set**	**number games**	R^2	MSE	null deviance	control deviance	model deviance
lifetime sales-to-date	pre-10	600	0.360	0.669	562	500	360
net first 10 week sales	pre-10	600	0.421	0.642	599	502	347
lifetime sales-to-date	meta	839	0.446	0.586	839	736	464
lifetime sales-to-date	meta sub	839	0.430	0.599	839	736	478

Table 1: Linear regression models. Review set indicates which subset of reviews were used: "pre-10" indicates only reviews from the 10 weeks prior to the launch of the game, "meta" indicates the full review database and metascore information, and "meta sub" is the same model without metascore information. Mean squared error (MSE) values are from leave-one-out cross-validation of models. Null deviance reports null model deviance; control deviance includes only console, game genre, and maturity (ESRB) rating as predictors; model deviance uses full set of predictive variables including controls.

view sentiment (mean vs median, variance), or extreme review sentiment (minimum and maximum).

Reviewer text analysis Our analysis of review text examined the predictive power of review text words to distinguish reviewers as professionals or amateurs (labeled 0 and 1, respectively) 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 from text analysis using the R programming language packages tm[6] and topicmodels[7]. 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.

For binomial regression we used R's glmnet package[8] 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. This text corpus was used to examine review texts but not for sales prediction tasks—future work should explore which text features are most predictive of game sales and how they relate to reviewer types.

Results

Before describing results on our prediction tasks it is important to understand the characteristics of the review and sales corpuses we collected. Metacritic amateurs 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 professionals whose reviews are featured on the site. Below we explore how professionals and amateurs differ in aggregate scoring and review sentiment expressions.

Review scores Amateurs and professionals show clear differences in review score assignment (Figure 1). Professionals tend to provide reviews distributed more tightly around a mean 7.2 and median 7.5 score, while amateurs show more variation with a mean 7.4 and median 9.0 score (Table 2). A Wilcoxon test found these differences to be significant ($p < 0.001$). Amateurs favor providing higher scores than professionals, but are the only ones likely to provide low review scores, seen by a larger portion of the cumulative review distribution on lower scores (Figure 1a). Amateurs tend to provide few reviews to any single game, while professionals tend to more frequently provide many reviews (Figure 1b). Thus, amateurs tend to focus on a few high-profile games, but provide a wider range of scores than professionals.

Factors relating to professionals' role and amateurs' motivations may explain these differences. Professionals are often under threat of blacklisting 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 professionals are unlikely to review low profile and low quality titles that will not drive traffic to their websites. These factors combine to skew professionals towards reviewing games generally favorably without providing overly positive reviews.

Amateurs, by contrast, are most likely to review provided a great or terrible experience (Gilbert and Karahalios 2010). 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. "Fanboy" culture also likely plays a role in driving reviewing behavior. Our text analysis results corroborate these findings, showing amateurs often describe product value while professionals focus on potential purchaser game feature interests and demographics (see below). Both amateur and professional reviews acknowledge the strongly split gaming demographics, shown through the prevalence of related terms (Table 5).

references? where should this go?

While previous studies have demonstrated both the bimodality and heavy-tailed distribution of number of reviews from amateurs, our results suggest professionals may have different habits and practice, suggesting different underly-

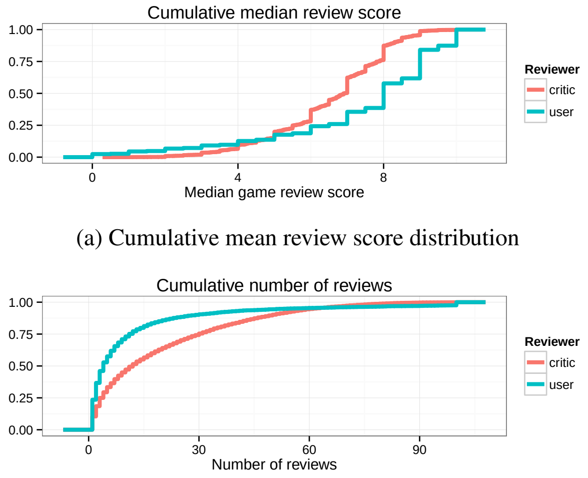
[6] http://cran.r-project.org/web/packages/tm/

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

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

	professionals		amateurs	
review attribute	**mean**	**median**	**mean**	**median**
score	7.18	7.5	7.44	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
min polarity	0.04	0.00	-0.17	-0.06
variance polarity	0.03	0.00	0.08	0.06
overall polarity	0.10	0.09	0.07	0.06
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
min subjectivity	0.36	0.37	0.15	0.00
variance subjectivity	0.04	0.00	0.10	0.11
overall subjectivity	0.59	0.60	0.61	0.60

Table 2: Comparison of professional and amateur review text sentiment.



(a) Cumulative mean review score distribution

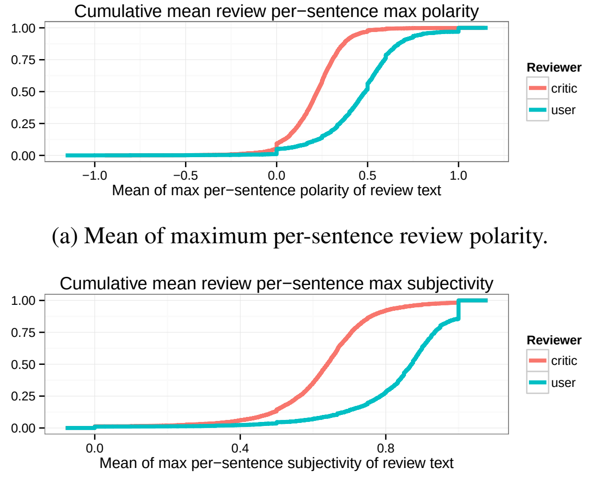
(b) Cumulative distribution of number of reviews per game

Figure 1: Game review corpus characteristics.

ing drives and purposes for reviewing that merit additional investigation.

Review sentiment Reviewers and professionals show similar average levels of review sentiment, but differ in expressions of extreme sentiment (Figure 2 and Table 2). For our analysis we considered both mean and median aggregations of polarity and subjectivity features over all reviews; all results reported were significant according to a Wilcoxon test at $p < 0.001$, although these differences may not appear in Table 2 when rounding to two decimal places. While differences in mean and median per-review polarity and subjectivity were significant, their magnitude was small. Both professionals' and amateurs' review distribution are slightly skewed toward more positive polarity and mild subjectivity (compare means and medians in Table 2).

Professionals and amateurs barely differ in mean polarity over a review, both centering on a mildly positive typical review. However, amateurs vary their sentence-level polarity more within reviews and show a strong skew towards at least one strongly positive and/or negative review sentence when compared to professionals (Figure 2a). Similarly, while professionals and amateurs show nearly the same level of moderate mean subjectivity, amateurs tend to use at least one highly subjective sentence (Figure 2b). Interestingly, amateurs are also more likely to employ at least one highly *objective* sentence, relating to their overall larger variance in expressions of subjectivity. Contrary to our expectations, professionals are not particularly objective when compared to amateurs, except in terms of the most polarizing and subjective sentences used in their reviews.



(a) Mean of maximum per-sentence review polarity.

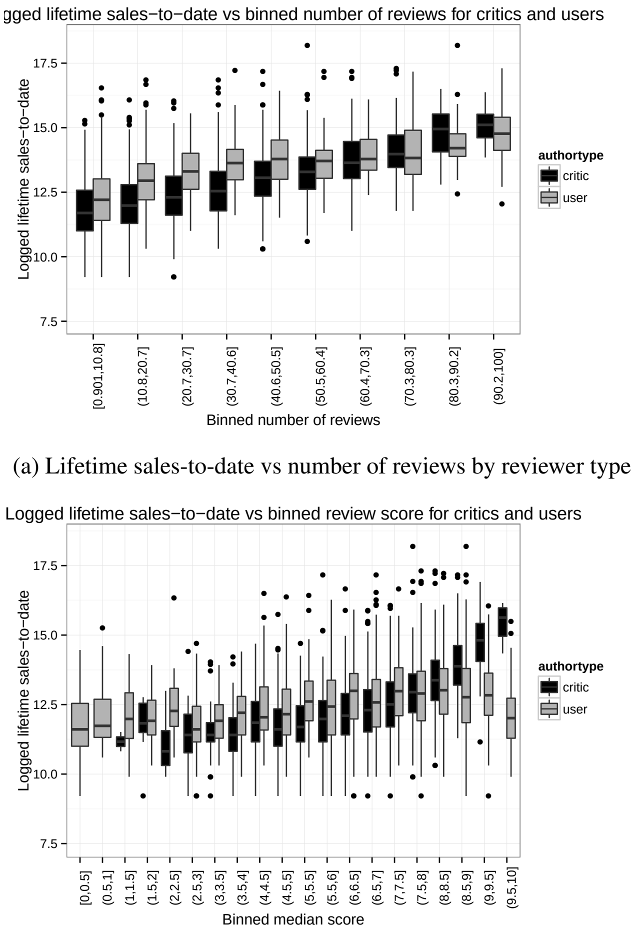
(b) Mean of maximum per-sentence review subjectivity.

Figure 2: Cumulative review sentiment (polarity and subjectivity) distributions comparing reviewer type

One limitation of this interpretation, however, is that professional 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 linear models to predict mean lifetime sales for each game, controlling for game console, genre, and Entertainment Software Rating Board (ESRB) maturity rating. While linear models may have limited predictive power for nonlinear relationships they afford direct interpretation useful to analysts employing the results. We examined models for predicting game sales-to-date and net first ten week sales considering a set of factors: review author (professional or amateur), review scores, number of reviews, and review text features (length, polarity, subjectivity). We examined four models: predicting total sales from reviews in the 10 weeks prior to the game's release (total vs pre-10), predicting net first 10 week sales from the same pre-release review sub-



(a) Lifetime sales-to-date vs number of reviews by reviewer type

(b) Lifetime sales-to-date vs mean review score by reviewer type

Figure 3: Mean lifetime sales-to-date vs review scores and volume.

set (10 week vs pre-10), and predicting total sales using all reviews available with metascore, professional, and amateur score information (total vs meta). Table 1 summarizes the models; Tables 3 and 4 summarize the predictive variables, standardized β coefficients, and model fitting results for each model.

Review scores and volume both show clear relationships to the lifetime sales of a game (Figures 3a and 3b). Anecdotally, many have claimed that a 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 professional scores, although the range of relevance is broader.[9] Amateur scores do not show a clear relationship to sales: increasing mean amateur review score does not appear to relate to any substantial gain in scores. By contrast, professional 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 professional 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.

Prediction Models Our models show moderately strong predictive power (R^2 between 0.360 and 0.446, MSE between 0.669 and 0.586) for logged sales (Tables 3 and 4). All predictive variables 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; "n.s." indicates a given coefficient was not significant. In all cases test models had a statistically significantly (χ^2 test, $p < 0.001$) smaller deviance than control models that accounted for game console, game genre, and game ESRB rating. Thus, all models explain variance in game sales not accounted for game market factors, showing the predictive values of Metacritic amateur and professional reviews.

For each predictive task we computed all possible linear models formed from subsets of review score and text features and scored them according to the Bayesian Information Criterion (BIC) using R's leaps package.[10] Using the BIC allows our model selection process to penalize overly complex models to reduce the likelihood of overfitting. Relative importance of model features was measured using the lmg metric in R's relaimpo package.[11] The lmg metric calculates the average R^2 contribution of a feature averaged over orderings among predictive variables—we use relative values among predictors to compare their importance.

Total and net 10 week sales vs pre-release reviews Marketers often seek to predict sales for unreleased games and must rely solely on information available prior to a game's release. For this application 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 both lifetime sales-to-date and net sales over the first 10 weeks after a game releases. The BIC identified the following predictive features: median review score, number of reviews, type of review (amateur or professional), and interactions between type of reviewer and median score, number of reviews, mean of mean per-sentence review polarity and mean of mean per-sentence review subjectivity. With an adjusted R^2 of 0.360 and MSE of 0.669 the pre-release review model moderately effective at predicting lifetime sales. Table 1 shows the slightly better performance of the pre-release review model when predicting first 10 week sales. Thus, pre-release reviews seem slightly more indicative of near-term sales than long-term sales. 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. In this dataset net first 10 week sales and total lifetime sales-to-date have a pearson's correlation coefficient of 0.861, confirming this relationship.

We compared these models with a third model that employed all reviews (rather than solely pre-release reviews)

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

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

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

for predicting total sales (“total vs all”); this model covered 2902 games, rather than the 600 in the previous models. This model gives an upper bound on performance for sales prediction given the information in our reviews, at least as we have processed it. All three models for sales prediction considering review features and their interactions with reviewer types show qualitatively similar effects (Table 3).

In general, number of reviews is most important for sales prediction, followed by median review score, and the interaction between median review score and reviewer type. Review sentiment factors are less important and have smaller coefficients in most cases. These three aspects account for most of these models’ predictive power. Median review scores have the most positive effect on sales with the interaction of these scores with reviewers as amateurs being equally powerful but *negative*. Higher scores suggest strong sales, but are not particularly important when coming from amateurs who tend to provide the same scores across games of varying commercial success. Number of reviews has a weaker positive effect than review scores. Together these results show professional scores and review volume indicate positive critical reception (and potentially game developer power), while amateur reviews are only useful as a gauge of popular interest.

more integration or potentially cut—more of a “discussion” than results

Sheer volume of amateur and professional reviews account for the largest part of our model’s power. Intuitively this makes sense: volume of professional and amateur reviews both reflect relative attention to a game and also likely reflect the relative publishing and marketing resources behind the game’s developer. More amateur reviews reflect games more people have played and thus likely already purchased. Further, games that drive amateurs to respond on online forums reflect greater social popularity. A greater volume of professional reception similarly reflects publisher budget and expectations. Review copies are typically distributed to professionals 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 professional scores and number of reviews have roughly equal relative importance, amateur scores are relatively unimportant. Given the distributions seen above this makes sense: games that have many amateur reviews are rare, but enjoy strong sales on average. For a game to have many amateur reviews requires many amateurs 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 amateur 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 amateurs or professionals devote to a game is indicative of its level of sales success. Future work should explore more sophisticated features for predicting sales such as volume of previews, media campaign efforts, and user interested expressed on Twitter, Facebook, Youtube and other media outlets. These sources may provide complimentary information on more broad interest in a product compared to those devoted enough to use Metacritic as a review outlet.

Comparing Metascores, professionals, and amateurs
To better understand the efficacy of different scoring systems we compared the predictive power of metascores, combined professional review scores, and combined amateur scores for predicting lifetime sales-to-date. From the total dataset, 840 games have data on lifetime sales-to-date, metascores, and reviews from both professionals and reviews. We predict lifetime sales as metascores are running values that are updated without a historical trace—thus we cannot know historical values of metascores prior to a game’s launch. Professional and amateur reviews were aggregated across the full set of data for a fair comparison with metascores.

According to the BIC criteria the best linear model—with an adjusted $R^2 = 0.4457$ and MSE $= 0.586$—uses (standardized β in parenthesis): metascores (0.449), median professional scores (-0.093), median amateur scores (-0.190), and number of professional reviews (0.282) (Table 4). Removing metascore information shifts the model to positively weight amateur review scores (0.223) and add number of amateur reviews (0.235) as a significant predictive factor. The model without metascores is a somewhat worse fit—$R^2 = 0.4298$ and MSE $= 0.599$ (χ^2 test $p < 0.001$).

As in the previous cases, volume of amateur and professional reviews captures a large portion of the model’s predictive power, using metascores for score information. The small differences in model fit suggest metascores are no more predictive than amateur or professional scores combined by simple weighting schemes (mean or median). That amateur scores were not predictive of sales when coupled to metascores surprised us. To understand this effect we examined the pearson’s correlation between metascores, amateur scores, and professional scores. Surprisingly, metascores are very strongly correlated with median amateur scores ($\rho = 0.961$) and more weakly correlated with median professional reviews ($\rho = 0.573$). The comparable overall model power between using only amateur and professional data for lifetime sales and adding metascores suggests the metascore weighting scheme may not be particularly powerful compared to simpler methods.

As a further comparison of reviewer efficacy for predicting sales success we computed the pearson’s correlation coefficient between the reviews given by a single reviewer and the logged total sales of all games that reviewer reviewed. We consider only reviewers with at least 10 reviews to avoid outliers, leaving 860 reviewers total. Examining the correlation distribution, we found 25% of reviewers have a negative correlation with total sales and that 16% of reviewers have a larger positive correlation than metascores ($\rho = 0.405$), and 22% of reviewers have a larger magnitude (positive or negative) correlation than metascores. Of those with larger magnitude coefficients, 82% were amatuers and 18% were professionals. Metascores are thus a good proxy for sales,

	10 week vs pre-10		**total vs pre-10**		**total vs all**	
variable	β	relative importance	β	relative importance	β	relative importance
median review score	0.64691	18.63%	0.553424	19.88%	0.30145	15.80%
number of reviews	0.23578	21.06%	0.235556	20.88%	0.53365	53.32%
reviewer type (as amateur)	0.43711	2.60%	0.606100	10.76%	0.37623	4.79%
reviewer type (as amateur) X median score	-0.61708	12.93%	-0.505708	11.01%	-0.21179	2.60%
reviewer type (as amateur) X number of reviews	0.51815	1.98%	0.431584	1.76%	n.s.	n.s.
reviewer type (as amateur) X mean(mean review polarity)	0.25741	4.70%	0.213716	4.50%	0.03635	2.31%
reviewer type (as amateur) X mean(mean review subjectivity)	-0.20898	2.77%	-0.170682	2.33%	-0.02785	0.17%

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

	with metascores		**no metascores**	
variable	β	rel. imp.	β	rel. imp.
metascore	0.449	11.49%	-	-
median professional score	-0.093	1.87%	-0.071	1.90%
median amateur score	-0.190	8.51%	0.223	15.68%
number reviews in metascore	n.s.	12.51%	-	-
number professional reviews	0.282	21.60%	0.280	28.28%
number amateur reviews	n.s.	12.62%	0.235	22.48%

Table 4: Regression model for lifetime sales-to-date comparing metascores, amateurs, and professionals. β are standardized regression coefficients when predicting log-scaled and standardized lifetime sales. All values are not listed as not significant ("n.s.") are significant at $p < 0.05$. Model $R^2 = 0.4457$, $MSE = 0.586$ when including Metascores, $R^2 = 0.4298$, $MSE = 0.599$ without. Metascore model is significantly different from null model (residual deviance 465 on 816 degrees of freedom, null deviance 839 on 838 degrees of freedom, χ^2 test $p < 0.001$). Model without Metascore is also (residual deviance 478 on 818 degrees of freedom, null deviance 838 on 838 degrees of freedom, χ^2 test $p < 0.001$). χ^2 test finds two models significantly different at $p < 0.001$.

but can be outperformed by individual reviewers. Sometimes this requires by taking the opposite of the scoring from a reviewer in cases where they have negative correlations with sales.

Professionals with high correlations tend to be trade magazines focused on particular consoles: Xbox Evolved (0.565), Computer Games Online RO (0.543), Playstation: The Official Magazine (US) (0.427). Some low correlation reviewers were surprising: The New York Times (0.114), The A.V. Club (0.149), GamePro (0.198), IGN (0.209), and GameSpot (0.211). The relative renown and influence attributed to several of these review sources highlight limitations of our analysis. We do not differentiate review authors from a given professional source, losing any predictive signal from publications with many different reviews (e.g., IGN, GamePro, GameSpot). Several of the low correlation sources are targeted towards small niche audiences among game consumers (e.g., The New York Times, The Onion's The A.V. Club). In addition, reviews are intended not only to predict sales, but can potentially act to dissuade others from game purchases or provide a thorough analysis. Future work should explore more nuanced aspects of how review aggregation platforms score venues along with automated techniques for weighting review source scores for sales prediction or other tasks.

Review Text Classification

Our previous analyses show reviews are predictive of sales. 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 professionals or amateurs using a penalized binomial regression model. Understanding these text-level differences can enable better prediction of what characteristics distinguish professional reviews in content from amateur reviews and how these text-level differences express general trends in how professionals and amateurs review.

The model achieved an F1 score of 0.9359 (precision 90.96% and recall 96.38%) over the set of 197,383 reviews we had data for. 96.38% of professionals were accurately labeled, and 77.28% of amateurs were correctly labeled, reflecting the imbalanced proportions of these categories in our data set—70.34% of reviews come from professionals, 29.66% from amateurs. 10-fold cross-validation found the model had a MSE of 0.4980. The penalized regression model employed 2,570 terms, assigning 146 a weight of zero, 1,235 positive weights (being an amateur), and 1,190 negative weights (being a professional). Below we examine several of the terms from these groups to understand how amateurs and professionals differentiate themselves in text descriptions with term coefficients in parenthesis. As words were stemmed we included completions of the stems in parentheses to help interpretation. Table 5 illustrates a selection of words with strong predictive power—β coefficients are reported in parentheses with positive values predicting a reviewer to be a amateur. Word categories were derived through inspection and are not part of the predictive model, but shown to highlight trends in how amateurs and professionals word their reviews.

Professionals were most recognizable by references to major game genres and the games market. Compared to amateurs, professionals 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 genres included descriptions of gameplay perspective—"firstperson" (-0.86) and "thirdperson" (-

1.01)—as well as gameplay style–“action” (0.36), “role-play” (-1.34), and “brawler” (-0.78). Professional reviews make an effort to characterize a game within the broader industry sales cycle and specific genres of interest. As professional reviews are often a form of marketing for companies this makes intuitive sense—professionals are acting to define the intended audience of a game through describing how it relates to a broader marketing strategy and genre. Professionals more often refer to game hardware features and the demographics of players who may be interested for these reasons.

Amateurs were recognizable by slang terms, references to game purchasing and reviewing, family, multiplayer, and aspects of the running game. Slang terms included phrases such as “wtf” (2.32), “imo” (4.59), “lol” (2.77), and “meh” (1.20). Reviews often mentioned aspects of reviewing games through references to review websites and aspects of “bias” (2.45) and “opinion” (1.46). Game purchasing appeared through purchase behavior, game pricing and purchase options, and places to purchase games (e.g., the “Gamestop” (3.31) retail chain). Amateurs also tend to reference the social dynamics of play, mentioning family and multiplayer behavior as well as general problems with actual game bugs or flaws. Mentioning live gameplay features and family members reflects amateurs’ stronger emphasis on the experience of playing with others, rather than assessing a potential product for (implicitly individual) consumption as done in professional reviews.

Combined, these results paint a picture of professionals taking a role of identifying purchase products and describing their features, while amateurs relate their games to broader consumption practices and gaming culture. Amateurs are free to reference particular game distributors or review biases in ways professionals cannot. Compared to professionals, amateurs typically relate games back to playing behavior, experiences, and social interactions. In part these results help understand the power of professional review scores to predict sales better than amateur review scores. Professionals describe games in a way to guide purchasing decisions, while amateurs 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

Our analyses of game sales and reviews uncovered aspects of the different roles amateurs and professionals play in promoting and consuming games. Professionals 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. In sum, professionals act as dispassionate professional sources of information and indicators of overall investment in game production values and marketing push. In contrast, amateurs 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 amateur reviewers as consumers evaluating the quality of a game experience and acting as indicators of overall market uptake and attention to the game.

Limitations

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 sales. Linear models allow for direct interpretation but lose the nuances of these almost certainly non-linear relationships. Further, 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 efforts, developer and publisher renown, and other external factors around sales regulation and economic circumstances all impinge on game sales. In addition, we are limited to third-party data and thus expect sales numbers to be approximate at best. That our models have as strong of correlations as they do is impressive in light of 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 review corpus is derived from a single website and professional reviews were limited to summary text. It is unclear whether the amateur reviewing patterns on other websites would show similar relationships as Metacritic is considered a prominent source for game review information. Perhaps the number of amateur reviews only matters from major review outlets, while other venues have weaker relationships with sales. Limiting professional text to summaries may have hurt the power of our models to detect any language professionals employ in a full review text. Yet this is typically what website users would find when seeking professional reviews. By employing summaries we maintained the interface of Metacritic, but this comes at the cost of missing the more general patterns of how professionals review when presenting materials on their own venue.

Our text-based models are limited to unigrams in a bag-of-words model. We lose information on linguistic qualifications, negations, and context that may provide additional nuance to reviews. The terms we found are thus an initial foray into the space of how reviews differ between professionals and amateurs. Even capturing simple adjective-noun relations would add depth and potential insight into what aspects of games reviewers describe and how they describe them (Zagal, Tomuro, and Shepitsen 2011).

Applications

The most obvious application for our work is predicting game sales from Metacritic reviews. Publishing firms can use this as a guide to predict sales based on trends in pre-release reviewing from professionals and amateurs. Such predictions can guide decisions to vary the resources devoted to marketing, developing, or supporting a game (particularly those with online or ongoing components). Marketers might attempt to drive amateur 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. Game marketers and producers could potentially quantify the value of reviews they get when selecting potential reviewers. Reviewers could develop novel ratings systems that are automatically scaled against other systems. Losing familiar amateur guides to scoring may hurt interpretability for amateurs, but may also allow both amateurs and professionals to focus more on aspects of game features or experience without filtering these results through an arbitrary number.

Conclusions

We presented an analysis of amateur and professional reviews and examined their power to predict game sales. Prerelease reviews can predict both early and long-term sales. Volume of response from both amateurs and professionals is most important, followed by professional (but not amateur) scores. Professionals emphasize game market niche while amateurs focus on gameplay experiences.

Acknowledgments

To be filled with non-anonymous information.

References

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

[Dellarocas, Zhang, and Awad 2007] 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, Gu, and Whinston 2008] 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 2011] Ehrenfeld, S. 2011. Predicting video game sales using an analysis of internet message board discussions. Master's thesis, San Diego State University.

[Ghose and Ipeirotis 2007] 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.

[Gilbert and Karahalios 2010] Gilbert, E., and Karahalios, K. 2010. Understanding deja reviewers. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 225–228. ACM.

[Gruhl et al. 2005] 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 2006] Liu, Y. 2006. Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing* 70(3):74–89.

[Marcoux and Selouani 2009] 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 et al. 2012] 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.

[Zagal, Tomuro, and Shepitsen 2011] Zagal, J. P.; Tomuro, N.; and Shepitsen, A. 2011. Natural language processing for games studies research. *Journal of Simulation & Gaming (S&G), Special Issue on Games Research Methods* 43(3):353–370.

category	word	β
game genre	roleplay	-1.34
	thirdperson	-1.01
	firstperson	-0.86
	brawler	-0.78
	openworld	-0.68
	singleplay(er)	-0.67
	oldschool	-0.47
	sidescrol(ler)	-0.32
game market	tie(-)in	-1.53
	competitor	-1.44
	holiday	-1.09
	followup	-0.64
	brand	-0.59
	predecessor	-0.50
	publish	-0.41
	budget	-0.40
hardware	peripher(al)	-0.58
	touchscreen	-0.50
player demographics	diehard	-1.09
	enthusiast	-1.09
	gamer	-0.32
	casual	-0.15
	fanboy	1.30
slang	awsom(e)	5.45
	imo	4.59
	lol	2.77
	wtf	2.32
game reviewing	metacrit(ic)	3.03
	ign	2.01
	review	1.74
	bias	2.45
	opinion	1.46
	critic	1.27
game purchasing	gamestop	3.31
	bought	1.92
	dlc	1.16
	preorder	1.06
	paid	0.75
	free	0.51
family	wife	2.74
	son	2.06
running game	beta	1.65
	server	1.40
	lagg(y)	1.27
	glitch	0.76
	glitch(y)	0.75
	bugg(y)	0.55
multiplayer	teamwork	0.57
	team	0.46
	splitscreen	0.43

Table 5: Major categories of review words distinguishing reviewers and professionals. Values in parentheses report β coefficients from binomial regression model where positive weights predict amateurs (as opposed to professionals).