Buying Into the Hype: Online Text Analysis for Video Game Sales Performance

|  |  |
| --- | --- |
| 000  00  00  00  00  00  00  00  00  00  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |

|  |
| --- |
|  |
| Noah Gale and Johannes Zhou |
|  |

Abstract

The increasing availability of user submitted content on the internet has provided an incredible source of data to observe and analyze populations of people. Methods of retrieval and analysis of this information is a well-established point of research. The large amount of comments that are posted daily to discussion boards and other communities can be invaluable in studying customers for business intelligence. In this study we utilize text data in the form of user submitted messages retrieved from popular online communities such as Reddit to model video game sales performance. We will use data mining techniques to gauge the level of enthusiasm customers have before a game’s release. We can test the predictive power of this model by comparing our results to existing sales data of those games. If methods to analyze comments continue to be developed and improved, these models allow many industries to predict the success of their products.

Introduction

The sheer amount of messages posted every second online in modern times makes analysis a daunting task. These challenges create a need for efficient automated methods to collect high volumes of this text based data and extract relevant information.  Current data mining methods give us the tools to dig into all the words in the comments and try a large variety of methods to synthesize a working model for video game sentiment. The primary goal of this study is to explore a precedent for creating these models. To analyze the comments we search for possible topic models and conduct sentiment analysis. On top of the text, we also look into the possible use of “likes”, or other forms of voting in each community, as an additional model feature. By mining features from comment text and using with available weekly sales data, we can see if sales performance can be predicted over time.

Related Work

Applying statistical analysis for business intelligence is not a very novel idea, and a number of researchers have already studied how online text could be used to make predictions in other industries. An early notable attempt was made with the paper *Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards* (Antweiler, & Frank, 2004). This study attempts to predict stock market movement from twitter chatter. While our goals are less ambitious, they are possibly more pragmatic. More relevant to our study, *Automated Text Analysis for Consumer Research* (A Humphreys, R Jen-Hui Wang, 2016) provides a review of existing methods of analyzing online text for consumer research. *The effect of social media communication on consumer perceptions of brands* (B Schivinski, D Dabrowski, 2014) looks at whether the perception of particular brands on Facebook have an effect on purchase intention. Finally, *Predicting video game sales using an analysis of internet message board discussion* (Ehrenfield, 2011) is almost perfectly in line with our study and serves as some inspiration for our research. The Ehrenfield study makes use of internet message boards and vgchartz sales data, but only uses the volume of discussion for particular games as a way to predict sales. We hope to build on this and add sentiment and text mining methods to the analysis. This can hopefully create a more accurate model and allow future models to gain more information from online text data.

Methods and Contribution

The problem for this study has two parts: data collection and analysis. Many studies focus on applying machine learning methods to existing datasets. We took steps to acquire the raw data and sort it into a usable dataset for analysis.

Data Retrieval

The first goal is to find relevant sources of data for video game discussion. The analytical challenge for this study is finding an effective method for extracting relevant information from online content. Online communities such as Reddit, Twitter, and Youtube all have comments related to video games, and are candidates for analysis. For this study we focused on the Reddit community and used RedditExtractoR to scrape comments from the site. These methods allow us to identify particular comment threads that are relevant to our study from the innumerable amount of other discussions on each site.

Sales data is also difficult to procure, as most of it is proprietary data. We will primarily use publicly available historical sales data on vgchartz.com, and try to look for more data or alternatives elsewhere.

For this study we examined game releases from 2014, 2015 and 2016, and are taking comments and sale data near release dates. Comment data is organized by thread and also sorted by date. In 470 thousand comments discussing 253 games. The sales data is recorded weekly over the entire year, providing total sales, publisher, and number of weeks since release for each game.

Before analysis we aggregated all raw data into one large dataset. Each comment was tagged with all relevant information – Reddit side information included the subreddit, user, time of post, and comment structure. Each comment was labeled with its corresponding game and relevant sales data such as time since release and sales numbers. The aggregate dataset was then cleaned, removing punctuation and stop words from comment text.

Analysis

The aim of this study was to determine if Reddit comments could be meaningfully used for business intelligence, specifically for gauging video game sales performance. Our approach was holistic and we went down multiple paths to extract useful information from the text data. The main four steps we attempted are sentiment analysis, WLS and OLS, topic modeling, and k-means. All analysis was done through Python and R. Textblob and sklearn were prominent packages used for analysis among other python and R packages.

Analysis and Results

Sentiment Analysis and Log Regression

|  |  |
| --- | --- |
|  |  |

Logistic regression is used to classify binary outcomes and in this study we tried to predict the continued presence of video games on the top 30 sales chart week to week. The model we utilized was comment based, from results gained from sentiment analysis.

Sentiments analysis divided comments by subjectivity and polarity. We used these outputs as features in this model, which factored in comment score, comment polarity, title polarity, and number of comments within a thread.

For sales position in week *t* we gathered our model’s information to predict the next week’s (*t+1*) outcome.

We were able to consistently reach an 80%-85% F1 score, the weighted measure of precision and recall, for predicting the following week’s performance for Week 1 to Week 4 on the top 30 chart. Notably, there’s no clear overall pattern to a video game’s performance in our observed sales dataset.

WLS and OLS

A simpler approach was to model sales numbers using weighted and ordinary least squares using the sentiment analysis output. Besides polarity and subjectivity values, we also considered comment scores, very much like with the logistic regression model.

This approach was found unsatisfactory, only reaching an accuracy of around 17%. Interestingly, when the model includes the publisher categorical variable, the one variable explained 60% of the variation.

Topic Modeling

Latent Direchlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) were used to procure topics from the text data.

Both approaches offered similar output, but topic modeling was a great challenge and difficult to implement for noisy, sparse text like internet comments, especially for a particularly casual and recreational subject. The topics derived from the output appear to offer no extra insight into the sentiments of the community users.

K-Means Analysis

Online community discussion boards like Reddit may have particularly influential users. Members that post more frequently, or generally are more recognized in the community may have a stronger effect on the overall sentiments of the community on subjects they comment on.

The K-Means model is an unsupervised approach that minimizes the within-cluster sum of squares. We tallied user post rate, upvote ratios, and mean comment scores to find possibly more influential users. We picked 10 clusters for this analysis.

We find that average upvote ratio across all users does not vary too much, mostly around the 80% range. Decent variance was observed for comment scores in all 10 clusters (fig 1). Mean ‘lateness’ and ‘activity’ were two variables we tried to add to the model. Lateness was the time between the original post and comment, and activity is the number of comments in said threads. Overall we find that there is noticeable heterogeneity between different clusters of Reddit users for all variables except upvote ratio. However, it was not possible to identify outstanding individuals from these variables.

Discussion and Future Goals

While text analysis continues to develop and models become more complex, internet chatter remains some of the most disorganized data to properly analyze. From the four approaches we tried, we continue to believe that there is potential in business prediction and market intelligence to be found within internet communities like Reddit, but the models required must be more robust. Better information retrieval is also an important step forward. An efficient way to collect and categorize the data we need did not exist, and would greatly improve the accuracy and scope of a study like this. Continued study of Reddit comments can consider the existence of more specialized subreddits. For our study, many dedicated subreddits for video games exist and may offer more varied, albeit biased discussion on the topics. Comment structure can also be examined. We approached each comment individually, and if replies and original posts could be differentiated, more information on the relationship between users can be used for analysis.

Acknowledgments

Special thanks to Professor David Jurgens for instruction and guidance for this project.

References

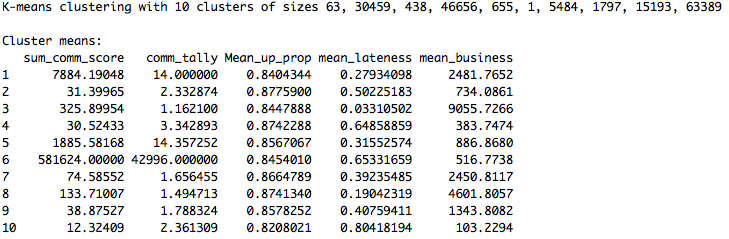
Antweiler, Werner, and Murray Z. Frank. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” *The Journal of Finance*, vol. 59, no. 3, 2004, pp. 1259–1294.,

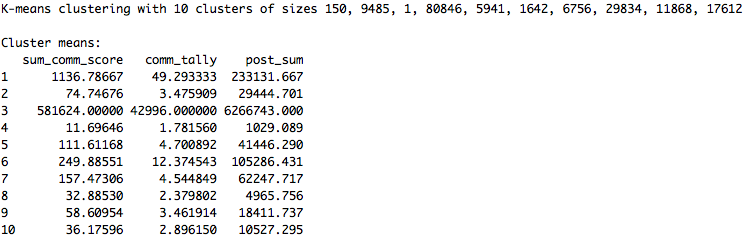
Humphreys, Ashlee, and Rebecca Jen-Hui Wang. “Automated Text Analysis for Consumer Research.” *Journal of Consumer Research*, 2017,

Schivinski, Bruno, and Dariusz Dabrowski. “The Effect of Social Media Communication on Consumer Perceptions of Brands.” *Journal of Marketing Communications*, vol. 22, no. 2, 2014, pp. 189–214.,

Ehrenfield, Steven. “Predicting Video Game Sales Using an Analysis of Internet Message Board Discussions.” San Diego State University, 2011

Appendix A: Figures

  
Fig1: K-means clustering output

  
Fig2: K-means clustering output (cont)