Can Tweets Kill a Movie? An Empirical Evaluation of the *Bruno* Effect

Omar Wasow
Harvard University
Barker Center, 2nd floor
Cambridge, MA 02138 USA
owasow@fas.harvard.edu

Alex Baron
abaron@post.harvard.edu

Marlon Gerra
mgerra@fas.harvard.edu

Katharine Lauderdale
klauderd@fas.harvard.edu

Han Zhang
zhang74@fas.harvard.edu

CHI 2010, April 10–15, 2010. Atlanta, Georgia, USA
ACM 978-1-60558-930-5/10/04.

Abstract

On Friday, July 10th 2009, the movie *Bruno* was number one at the box office and took in over $18.8 million in revenue. Based on this initial performance, analysts predicted the movie would rake in over $50 million in its opening weekend. By Saturday, however, the movie experienced an unusually sharp 38% decline in box office receipts. Several prominent journalists speculated that comments on the social media site Twitter.com may have amplified negative word-of-mouth about the movie and caused the dramatic fall-off in revenue. We investigate this "*Bruno* effect" and, contrary to popular accounts, find that neither positive nor negative comments on Twitter are associated with changes in box office performance. We do find suggestive evidence, however, that the volume of tweets, while not necessarily altering consumer behavior, may provide useful information for predicting opening weekend box office performance.

Keywords

Twitter, Movies, Word of Mouth, Box office

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User interfaces – Evaluation/ methodology

Introduction

Over the last decade, online social media has grown explosively and transformed how people communicate, transact, organize and are entertained. Twitter is a leading social media site in which users can post short, 140 character messages or "tweets" for public viewing. The site was founded in 2006 and is now the thirteenth most popular site on the Internet [1]. Over the last three years, Twitter has been credited with playing an important role in a wide variety of contexts including political campaigns, legal proceedings, citizen activism, news reporting and emergency response. Twitter has also been described as dramatically amplifying the effects of word-of-mouth feedback among consumers. In the summer of 2009, the movie *Bruno*, was released and was number one among movies released that weekend. Between Friday and Saturday, however, the movie experienced a 38% drop-off in revenue, substantially more than other new movies. Widespread speculation among journalists and bloggers suggested that word-of-mouth on twitter might have "killed *Bruno*" [3, 6, 8, 4]. The evidence for these assertions, however, was largely anecdotal.

This paper investigates if comments on Twitter exhibit any relationship to box office revenue over opening weekend. Echoing the questions posed by journalists and bloggers, we focus particularly on whether tweets on the Friday of opening weekend are associated with future changes in revenue, especially changes in the percent change in box office receipts between Friday and Saturday.

On a broader level, we look at whether Twitter microblogging for a film has any explanatory power for box office gross above existing models. We also investigate the relationship between a movie's Twitter presence and various other measures of its success (e.g. critical reception). More specifically, we investigate whether sentiment expressed on Twitter could plausibly influence movie box office performance in the way suggested by commentators around Bruno's opening weekend.

Literature Review

Standard predictive models of movie performance do not include word-of-mouth data; Basuroy, Chatterjee & Ravid (2003) concluded that critical reviews, budgets and star power are the strongest predictors of box office success[2]. Using weblog data, Mishne & Glance (2005) found that positive sentiment in posts is a better predictor of box office success than the volume of discussion alone [7]. Using IMDB movie information and the Blogpulse index, they concluded that positive sentiment in the pre-release period was better correlated with movie success than pre-release blog post count was. Zhang & Skiena (2009) used IMDB and movie news data to predict box office grosses, and found that the volume and sentiment of movie news coverage improved the predictive performance of a model based only on IMDB data (budget, number of screens, etc.) [9]. In addition, a market research firm conducted a small analysis of Twitter's influence on *Bruno*. Hampp (2009) reported on a comparison between Twitter traffic for *Bruno* and three other summer movies during their opening weekends [5]. Although it was found that *Bruno* had the highest number of negative tweets and negative percentage change between first- and second-day grosses, Hampp emphasized that the analysis cannot attribute causal influence to Twitter.

Data:

This project used three sets of data on Twitter and three on movie performance to address our questions of interest. The first Twitter dataset was approximately 200,000 downloaded tweets that each mentioned one of 58 different current movies. This data came from the website TwitCritics.com,

Table 1: **Summary Statistics, Movie and Twitter Data**

	Mean	SD
# Theaters on Day 1	1,744	1,307
Budget	49,094,810	4,5638,800
Mean Critics Reviews	0.45	0.24
Num Tweets Day 1 + 2	1,189	1,515
Day 1 Revenue	6,971,092	10,603,440
Day 2 Revenue	6,440,534	7,067,621
Day 2 Percent Change	0.11	0.29
Mean Sentiment Day 1	0.73	0.12
# Tweets Day 1	447	723

which aggregates tweets by their referenced movie and performs a sentiment analysis on each tweet, evaluating it as either a positive or a negative review. Table 1 presents summary statistics for the movie and twitter data.
TwitCritics.com only provides information for movies launched since fall of 2009, and so the second Twitter dataset, which produced the information for *Bruno*, came from Tweetscan.com, an archive of tweets. We then developed a sentiment analysis model to automatically predict sentiment ratings of tweets, and to replicate the algorithm provided by TwitCritics.com. [2]

The third set of Twitter data was used as a diagnostic for the TwitCritics sentiment analysis. The data came from Amazon Mechanical Turk (MTurk), a virtual marketplace that enables computer programs to co-ordinate the use of human intelligence to complete tasks that are difficult for computers to perform. We posted 2,787 of our TwitCritics tweets (approximately 100 per movie for 31 movies) to be evaluated as either extremely negative, negative, neutral, positive or extremely positive. Each tweet was evaluated individually by a person. We then had a human-coded sentiment to compare with the TwitCritics sentiment analysis. Table 2 presents a summary of the comparison between the human coded and algorithmic methods employed by TwitCritics.com. As a simple validation check, the MTurk data suggest that the TwitCritics ratings are fairly accurate, especially when we have a sufficient number of tweets. Further diagnostics follow later in the paper.

For movie performance data, we downloaded variables including box office total gross revenue, theaters, production budget, genre, and a proxy for actors' star power for each of the movies□ essentially all basic information that we could obtain and which prior literature suggested was poten-

[2]Using the data set of 200,000 tweets and ratings from TwitCritics, we then trained the content analysis model in two stages. First, noise words (like "the," "about," "from,") and noise characters (punctuation and non-ASCII characters) were removed from the text of the tweets. We then produced a set of statistics for each word in the training data that included: percentage of total occurrences that the word appeared in a positive statement, an indicator variable for whether the word only appeared in a positive or negative statement, and the total number of occurrences of the word in the training data. In the second stage, the model was supplemented with hundreds of known positive and negative words. Due to the high sentiment reliability of these words, the model was adjusted to weight them highly.

tially relevant for predicting box office performance. Most of this information came from the-numbers.com and boxofficemojo.com, movie data aggregators. Lastly, we downloaded data on critical reception by journalists from RottenTomatoes.com, which aggregates critical reviews for films, determines whether each movie is positive or negative, and then calculates the percentage of positive reviews for each movie (its "Tmeter").

Analysis:

We are interested in predicting the change in revenue between the first and second day of a movie's release. Our response variable is the percentage change in revenue (the difference in revenue between day one and day two divided by the day one revenue). We used two different measures of Twitter activity for each movie. The first measure was the number of tweets for each movie on the first day of release. The second variable was the mean sentiment (between 0 and 1) for each movie on the first day of release. We included these two different measures because they capture different aspects of the word-of-mouth presence, namely, the volume and positivity of Twitter attention.

Initial analyses showed that models with all movie performance variables included too many confounders to reveal meaningful relationships, and subsequently we investigated parsimonious models more appropriate for our questions of interest. In particular, the indicator variable we constructed for sequel was not significant. We also attempted to incorporate a variable for "star power," a measure of the box office success of the particular actors in each movie. Although RottenTomatoes.com provides a variable for this, it is calculated from the same set of critical review sentiments as the Tmeter variable, and to avoid multicollinearity, we omitted the star rating from the model.

After the preliminary data analysis exploration (not shown) we then constructed a full model including all variables that seemed significant and not obviously confounded. This model included the following covariates: the log of the number of tweets on day one, the mean tweet sentiment, the log of the production budget, the number of theaters on day one, and the Tmeter (the RottenTomatoes.com measure of critical reception).

Full Model:

$$\begin{aligned} \hat{y}_{\%ch})|(x) &= \beta_1 log(x_{\#tweets}) + \beta_2 x_{sentiment} + \\ &+ \beta_3 log(x_{budget}) + \beta_4 log(x_{thtrs}) + \\ &+ \beta_4 log(x_{tmeter}) + \epsilon \end{aligned} \quad (1)$$

Table 3 presents the regression results of the full model. In this model, two variables are significant, the budget and the number of number of tweets. We then constructed a reduced model with all insignificant variables removed:

Reduced Model:

$$\hat{y}_{\%ch}|(x) = \beta_1 log(x_{\#tweets}) + \beta_2 (x_{budget}) + \epsilon \quad (2)$$

Table 4 presents the regression results of the reduced model. We then perform an F-test to determine whether the full model has significantly greater predictive power than the reduced model. Our F-statistic was 0.11 and the corresponding p-value was 0.95. Our p-value is well above 0.05, so we cannot reject the null hypothesis that the full model and reduced model are indistinguishably predictive.

The significant correlation between revenue and budget was expected and consistent with the literature. However, the different findings for the two twitter variables, although

Table 2: **TwitCritics and Human Coded Sentiment Ratings, Select Movies**

	Movie Name	# Tweets	MTurk Sent.	Twit. Sent.	Diff.
1	Amreeka	6	1.00	1.00	0.00
2	Extract	75	0.79	0.69	0.09
3	Gamer	78	0.71	0.74	-0.04
4	Sorority Row	79	0.86	0.82	0.04
5	Tyler Perry's I ...	92	0.95	0.98	-0.03
6	Whiteout	74	0.36	0.39	-0.03
7	Bright Star	20	0.85	0.70	0.15
8	Burning Plain	19	0.79	0.79	0.00
9	Cloudy With a ...	91	0.99	0.96	0.03
10	Informant!	70	0.83	0.73	0.10
11	Jennifers Body	77	0.74	0.69	0.05
12	Love Happens	82	0.79	0.78	0.01
13	Fame	71	0.83	0.83	0.00
14	Pandorum	75	0.87	0.83	0.04
15	Surrogates	71	0.72	0.69	0.03
16	Invention of ...	78	0.79	0.78	0.01
17	Zombieland	85	0.98	0.93	0.05
18	Serious Man	21	0.90	0.67	0.24
19	Couples Retreat	87	0.92	0.91	0.01
20	Law Abiding ...	90	0.94	0.92	0.02
21	The Stepfather	75	0.75	0.71	0.04
22	Where The Wild ...	79	0.90	0.84	0.06
23	Amelia	50	0.60	0.66	-0.06
24	Astro Boy	73	0.92	0.84	0.08
25	Saw VI	83	0.83	0.69	0.14
26	Cirque du Freak: ...	82	0.93	0.89	0.04
27	Michael Jackson's ...	80	0.96	0.90	0.06
28	The Box	78	0.46	0.49	-0.03
29	Disneys A Christmas ...	87	0.92	0.87	0.05
30	The Fourth Kind	67	0.79	0.82	-0.03
31	The Men Who Stare ...	82	0.77	0.78	-0.01

counterintuitive, partly corresponds to prior articles in which valence is less predictive than volume. A scatterplot of the relationship between the mean sentiment of tweets and the percent change in revenue between Day 1 and Day 2 can be seen in Figure 1.

While the effect of the positivity of tweets was not significant, the effect of the number of tweets was statistically significant but negative (both the sign and significance were robust to a wide variety of model specifications). From the results in Table 4, we can back-transform coefficient on the logged number of tweets and, holding budget constant, find that a

Table 3: **Regression Results, Full Model**

	Estimate	SE	t-value	Pr($>\|$t$\|$)
(Intercept)	-0.13	0.91	-0.14	0.89
log(# Tweets Day 1)	-0.09	0.03	-3.07	0.00
Sentiment Day 1	0.16	0.31	0.53	0.60
log(Budget)	0.04	0.05	0.70	0.49
# Theaters	-0.00	0.00	-0.24	0.81
$\sqrt{Critics}$	-0.07	0.20	-0.33	0.74

Table 4: **Regression Results, Reduced Model**

	Estimate	SE	t-value	Pr($>\|$t$\|$)
(Intercept)	0.01	0.77	0.02	0.99
log(# Tweets Day 1)	-0.09	0.03	-3.52	0.00
log(Budget)	0.03	0.04	0.75	0.45

doubling in the number of tweets is associated with a 0.91 decrease in the percentage change in revenue from day one to day two. A scatterplot of the relationship between the logged number of tweets and the percent change in revenue between Day 1 and Day 2 can be seen in Figure 2.

Diagnostics:

Diagnostic plots (not shown) suggest that the model assumptions hold; namely, the residuals appear independent and normally distributed. The Shapiro-Wilk test for normality of the residuals produces a test statistic of 0.97 with a corresponding p-value of 0.50. We cannot reject the null, and can conclude that the residuals are normally distributed. Plotting residuals against leverage, suggests the residuals all lie within an acceptable range of Cook's distance. Moreover, the model was robust to removing outlying observations.

In addition to the model diagnostics, we also evaluated the TwitCritics sentiment analysis by comparing a subsample of the exact same tweets to human coders on Amazon Mechanical Turk (MTurk). For each movie we took the average sentiment rating from MTurk and compared them to the average rating obtained from TwitCritics. Performing a paired t-test, the results show there is a statistically significant difference in the two sample means of 0.036 (on a 0-1 scale) with a two-sided p-value of 0.002. However the magnitude of this difference, for practical purposes, is quite small (see Table 2). Although the paired t-test shows that there may be some discrepancy between the TwitCritics and MTurk content analyses, in general the MTurk results appear to corroborate the sentiment ratings from TwitCritics.

Conclusions:

Contrary to popular accounts, we do not find a *Bruno* effect in which negative word-of-mouth on Twitter is associated

with a notable decline in opening weekend box office performance. We do, however, find that the number of tweets has a significant association with a negative change in revenue between Friday and Saturday. In short, volume appears to matter and valence does not. While it is possible that faulty sentiment analysis is responsible for the non-significance of the sentiment finding, our validation check via MTurk offers some indication that the sentiment ratings are plausible. An alternative interpretation is that the amount of discussion about a movie is a better indication of valence or sentiment than what people actually say. However, our explanations are speculative and require further data collection and analysis for verification.

Our data had limitations that narrowed the scope of our inference. First, most of our tweet data was censored, in that it was almost all mined on two dates. This meant that it was not possible to include data from much beyond opening weekend in the analysis, because the different movies had been in theaters for various and non-comparable periods of time. Although it would have been preferable, it was not possible to obtain historical data on movies that had completed their entire theater runs due to limitations in the Twitter and TwitCritics archives. The necessary discarding of the majority of our Twitter data was also not ideal, and a larger data set containing more movies, collected over a longer period of time, could improve the analysis. If the relationship between tweet volume and revenue during opening weekend is also present in larger and different data sets, then Twitter is potentially a valuable source of nearly free and real-time public opinion that could provide useful information for forecasting movie performance. With more data, a predictive relationship between pre-release tweeting and opening weekend performance could also be investigated.

References

[1] Traffic stats for twitter.com, January 2010.

[2] S. Basuroy, S. Chatterjee, and S. Ravid. How critical are critical reviews? the box office effects of film critics, star power, and budgets. *Journal of Marketing*, 67(October):103–117, Jan 2003.

[3] R. Corliss. Box-office weekend:Brüno a one-day wonder?, July 2009.

[4] J. V. Grove. Did opening night twitter reviews sink *Bruno's* weekend box office?, July 2009.

[5] A. Hampp. Forget Ebert: How Twitter makes or breaks movie marketing today, 2009.

[6] M. Keane. Did Twitter bury "Bruno'"?, July 2009.

[7] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[8] A. Salkever. Did Twitter kill 'Bruno'? maybe not, July 2009.

[9] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. *wi-iat*, 1(2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology):301–304, 2009.

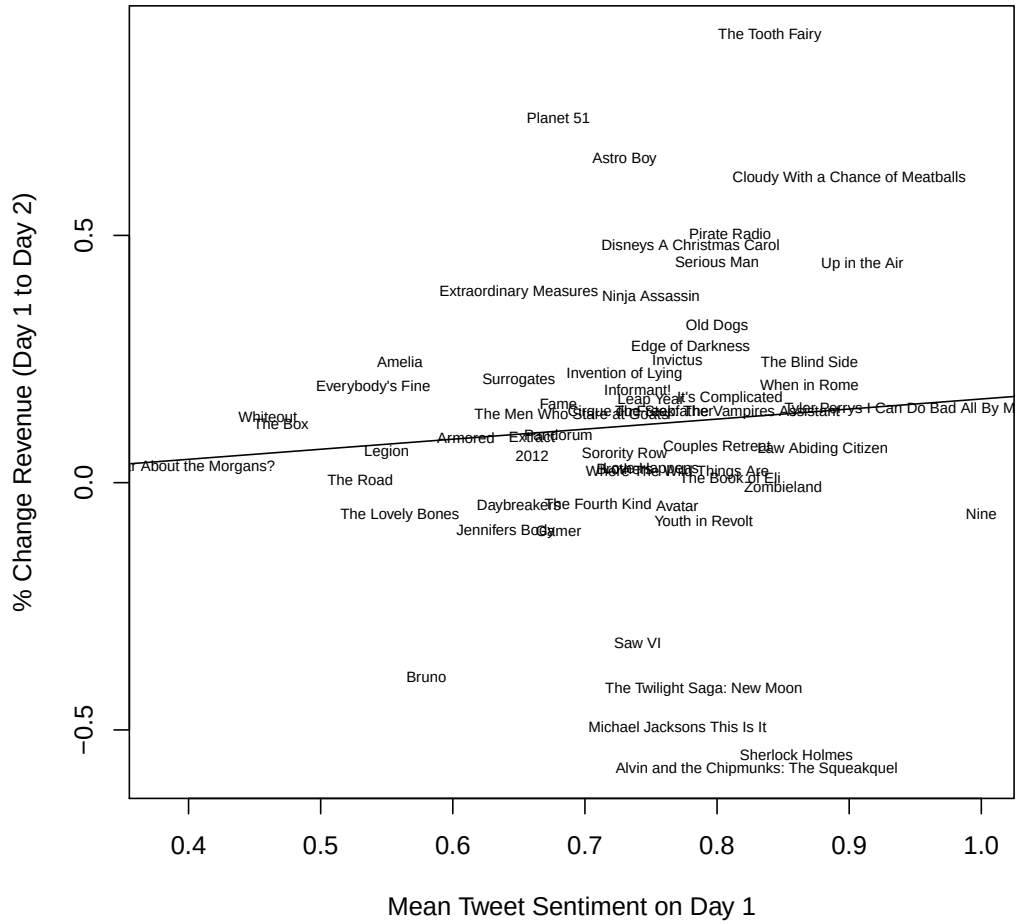

Figure 1: **Twitter Sentiment vs. % Change in Day 1 to Day 2 Revenue**

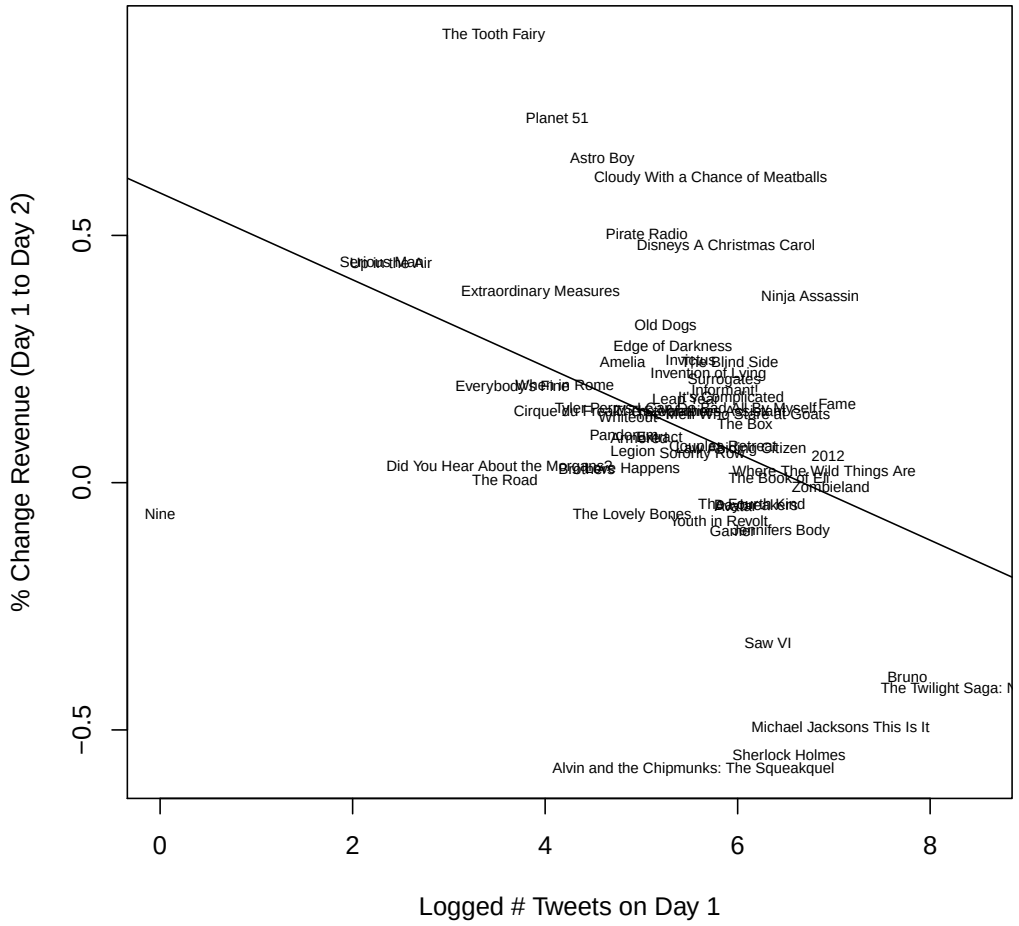

Figure 2: **Twitter Volume vs. % Change in Day 1 to Day 2 Revenue**