Predicting Movie Sales from Blogger Sentiment

Gilad Mishne

Informatics Institute, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad@science.uva.nl

Natalie Glance

Intelliseek Applied Research Center
5001 Baum Blvd, Pittsburgh, PA 15213
nglance@intelliseek.com

Abstract

The volume of discussion about a product in weblogs has recently been shown to correlate with the product's financial performance. In this paper, we study whether applying sentiment analysis methods to weblog data results in better correlation than volume only, in the domain of movies. Our main finding is that positive sentiment is indeed a better predictor for movie success when applied to a limited context around references to the movie in weblogs, posted prior to its release.

If my film makes one more person miserable, I've done my job.
– Woody Allen

Introduction

Weblogs provide online forums for discussion that record the voice of the public. Woven into this mass of discussion is a wide range of opinion and commentary about consumer products. This presents an opportunity for companies to understand and respond to the consumer by analyzing this unsolicited feedback.

Over the past year, marketing and public relations experts have tuned in to the channel of opinion output by online chatter. WOMMA, the Word of Mouth Marketing Association, as of early October 2005, boasts over 200 members. The influence of blogs comes in different forms. On the one hand, blogs are important because they represent the aggregate voice of millions of potential consumers. Admittedly, the demographics of bloggers is skewed towards high school and college age. In particular, earlier in the year, a Pew Internet and American Life study found that 48% of bloggers are under 30.[1] However, if blog data is sliced by demographic profile, then the skew can be factored out.

On the other hand, some individual bloggers have earned themselves a disproportionate share of influence, with readerships rivaling that of regional newspaper columns. As a result, commentary posted by an influential blogger often cross-polinates with editorial and news articles published in mainstream media. Recently, Jeff Jarvis posted about his negative experience with Dell computer support.[2] Although Jarvis' blog is mostly commentary about international news,

his unhappiness with Dell as a company reverberated across the blogosphere and into the press, creating a public relations mini-fiasco for Dell.

Hard evidence of the influence of online discussion on consumer decisions is beginning to emerge. An Intelliseek survey of 660 online consumers showed that people are 50 percent more likely to be influenced by word-of-mouth recommendations from their peers than by radio/TV ads [3]. Researchers at IBM reported that online blog postings can successfully predict spikes in the sales rank of books (Gruhl, Guha, Kumar, Novak, & Tomkins, 2005), showing that the raw number of posts about a book was a good predictor.

However, opinion comes in many flavors: positive, negative, mixed, and neutral mixed in with splashes of sarcasm, wit and irony. Novel techniques in sentiment analysis make it possible to quantify the aggregate level of positive vs. negative mentions with reasonable accuracy. (Although identifying the more subtle nuances of expression remains a challenge for machines as well as for many humans.)

The main question addressed in this paper is whether or not aggregate polarity scores are a better predictor of sales than simple buzz count. In particular, we analyze the sentiment expressed in weblogs towards movies both before the movie's release and after, and test whether this sentiment correlates with the movie's box office information better than a simple count of the number of references in weblogs does.

Related Work

Work on sentiment analysis of reviews and opinions about products is plentiful, particularly in the domain of movies. Typically, the methods employed include combinations of machine learning and shallow NLP methods, and achieve good accuracy (Dave, Lawrence, & Pennock, 2003, Liu, Hu, & Cheng, 2005, Pang, Lee, & Vaithyanathan, 2002, Turney, 2002). In particular, a recent study showed that peaks in references to books in weblogs are likely to be followed by peaks in their sales (Gruhl et al., 2005). Likewise, earlier work showed that references to movies in newsgroups were correleted with sales (Tong, 2001).

Our work differs from existing studies of sentiment analysis and business data in two important respects. First, our

---

[1] http://www.pewinternet.org/pdfs/PIP_blogging_data.pdf

[2] http://buzzmachine.com/archives/cat_dell.html

[3] http://intelliseek.com/releases2.asp?id=141

domain is weblog posts, a domain which tends to be far less focused and organized than the typical product review data targeted by sentiment analyzers, and consists predominantly of informal text. Second, we test how well sentiment performs as predictor for future sales information. All but one of the other studies cited examine correlation, not prediction. To this end, we examine sentiment in two separate time periods: both prior to, and after the product's release. Gruhl *et. al* also measure the predictive power of buzz on sales, but do not use sentiment analysis for their prediction experiments.

Data and Experiments

We now present the data and methodology we used for examining the correlation between sentiment and references in weblogs.

Business Data. We used IMDB—the Internet Movie Database—to obtain, for each movie, the date of its "opening weekend" (the first weekend in which the movie played in theaters), as well as the gross income during that weekend and the number of screens on which the movie premiered. We focus on the opening weekend data rather than total sales since this normalizes the figure across movies that were released on different dates, preventing earlier movies from having a higher income just because they have been "out there" longer, have been released on DVD, etc. Opening weekend income typically correlates highly with total movie income, accounting for an estimated 25% of the total sales (Simonoff & Sparrow, 2000). The number of screens the movie premiered on was used to normalize the opening weekend income, producing a "Income per Screen" figure for each movie. This allows comparing sales of summer blockbuster movies, sometimes released to 4000 screens in the opening weekend, to lower-profile movies released to 1000–2000 screens.

Weblog Data. For each movie, we collected all relevant weblog posts appearing in the Blogpulse (Glance, Hurst, & Tomokiyo, 2004) index. A post was considered "relevant" to a movie if the following conditions hold:

- The date of the post is within a window starting a month prior to the movie's opening weekend date and ending one month after it.

- The post contained a link to the movie's IMDB page, *or* the exact movie name appeared in the post in conjunction with one of the words ⟨movie, watch, see, film⟩ (and their morphological derivatives).[4]

For each relevant post, we extracted the contexts in which the movie was referenced by taking a window of k words around the hyperlinks to the movie IMDB page, or around

---

[4]Our strategy aimed for high precision without overly sacrificing recall. An examination of the retrieved posts shows a high degree of precision. We did not explicitly measure recall, but did find that using a combination of an IMDB link query and text queries increased recall by a factor of 500% over simply using the IMDB link query, which has near-perfect precision but limited recall.

exact matches of the movie names; we used varying values for k, from 6 words to 250. Finally, we calculated the sentiment value of the contexts using the methods described in (Nigam & Hurst, 2004).

Examples of different context lengths for the same reference to a movie are shown in Table 1, along with the computed polarity; this demonstrates the possible errors which result from taking "too much" or "too little" context.

In total, our analysis was carried out over 49 movies; these consist of all movies released between February and August 2005, with a minimum budget of $1M, and with publicly-available sales information. A sample item used in our experiments is shown in Table 2. Note that the polarity score is fitted to a log-linear distribution, with the majority of scores falling within a range of 4 to 7 (Nigam & Hurst, 2005). Thus, the average polarity score of 5.5 for the movie in the table indicates significant positive overall sentiment.

Movie	The Sisterhood of the Traveling Pants
Opening Weekend	5 June 2005
Opening Weekend Sales	$9.8M
Opening Weekend Screens	2583
Income per Screen	$3800
Pre-release Data	
References in weblogs	1773
Context Length: 10 words	
- Positive references	329
- Negative references	50
- Mean sentiment polarity	5.5 / 10
Context Length: 20 words	
. . .	. . .
Post-release Data	1618
. . .	. . .

Table 2: Sample data from our collection.

Experiments

We measured Pearson's r-correlation between several sentiment-derived metrics and both income per screen and raw sales. Among the sentiment metrics were: the number of positive contexts; the number of negative ones; the total number of non-neutral contexts; the ratio between positive and negative contexts; the mean and variance of the sentiment values. In addition to the sentiment-related correlations, we also measured the correlation to the raw counts of occurrences in weblogs (the number of weblog posts in which the movie is referenced). Measurement was done separately for pre-release contexts and post-release ones.

Raw counts vs. sentiment values. Our first observation is that usage of the sentiment polarity values, given the optimal context length, results in better correlation levels with movie business data than the raw counts themselves *for data gathered prior to the movie's release*. For data gathered after the movie's release, raw counts provided the best indicator. Of the different polarity-based measures used in our experiments, those yielding the best correlation values were as follows:

Length	Content	Sentiment
10	Rach went to see "The Sisterhood of the Traveling Pants" we both read the	Neutral
40	gym or work. 2. It was the first time I laughed since I've been home. Me and Rach went to see "The Sisterhood of the Traveling Pants", we both read the book and the girl from Gilmore Girls was in it. I had the best time	Positive
120	Tonight was a night of firsts. 1. I'm pretty sure it was the first time I left the house and didn't go to the gym or work. 2. It was the first time I laughed since I've been home. Me and Rach went to see "The Sisterhood of the Traveling Pants", we both read the book and the girl from Gilmore Girls was in it. I had the best time with her. We laughed, were both such dorks. The movie was SOOO sad. Like true "The Notebook" quality. I enjoyed it and it got me thinking. I need to stop being so miserable. I make my time here suck. I	Negative

Table 1: Polarity of different contexts.

- Prior to the movie release: the number of positive references within contexts of length 20.
- After the movie release: the number of non-neutral references within contexts of size 140 (using the number of positive references achieves very close results).

Table 3 compares the correlation between movie business data for raw counts and for the best performing polarity-related metrics. Clearly, the sentiment-based correlation improves substantially over the raw counts for pre-release data, whereas for post-release data the effect is negative (but minor).

Correlation	Between …	Period
0.454	Raw counts and income per screen	
0.509 (+12%)	Positive contexts and income per screen	Pre-release
0.484	Raw counts and sales	
0.542 (+12%)	Positive contexts and sales	
0.478	Raw counts and income per screen	
0.471 (-1%)	Non-neutral contexts and income per screen	Post-release
0.614	Raw counts and sales	
0.601 (-2%)	Non-neutral contexts and sales	

Table 3: Comparison of correlation between movie business data and blog references, with and without use of sentiment. Context sizes used: 20 (pre-release), 140 (post-release)

While the improvement using sentiment values on pre-release data is in-line with intuition, it is unclear to us why is does not have a similar effect for post-release data. One possible explanation is that post-release contexts are richer and more complex, decreasing the accuracy of the sentiment analysis.

Context Length. Our next observation is that constraining the context being analyzed to a relatively small number of words around the movie "anchor" is beneficial to the analysis of pre-release polarity metrics, but reduces the effectiveness of the post-release metrics. Figure 1 displays the relation between the correlation values and the context length for two particular instances of analysis: the correlation between the number of positive contexts before the movie release and the income per screen, and the correlation between the number of non-neutral contexts after the release and the opening weekend sales for the movie.

Examining the contexts extracted both before and after the movie's release, we note that references to movies be-
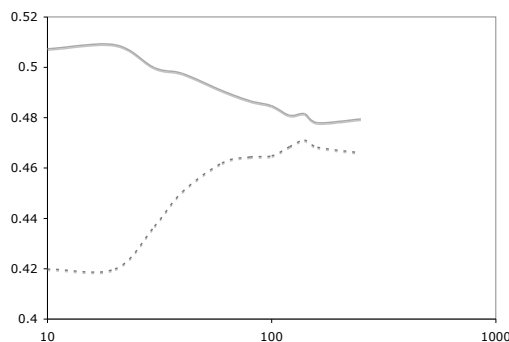

Figure 1: Relation between context length and correlation to income per screen: positive references, pre-release (blue, continuous line) and non-neutral references, post-release (red, dashed line). The X-axis shows the context length, and the Y-axis shows the level of correlation.

fore their release tend to be relatively short, as the blogger typically does not have a lot of information about the movie; usually, there is a statement of interest in watching (or skipping) the movie, and possibly a reaction to a movie trailer. References to movies after their release are more often accounts of the blogger's experience watching the movie, containing more detailed information – see an example in Table 4. We hypothesize that this may be the explanation for the different effect of context length on the correlation quality.

Breakdown

Figure 2 shows the degree of correlation between income per screen and positive contexts, both normalized across all movies.

Out of the 49 movies in our study, over half have very good correlation between pre-release positive sentiment and sales. Less than 20% can be viewed as outliers: movies whose average income per screen was poorly predicted by pre-release sentiment. How can the low correlation between weblog opinion and business data be explained for these outliers? In fact, movie sales have been shown to be affected by many factors unrelated to online discussion, such as genre, Motion Picture Association of America rating, other movies released at the same time, and so on (Simonoff & Sparrow, 2000). On top of that, noise originating from different com-

apparently an early easter is bad for apparel sales. who knew? i'll probably go see "guess who?" this weekend. i liked miss congeniality but the sequel [link to IMDB's page for "Miss Congeniality 2"] looks *awful*. and seattle's too much of a backwater to be showing D.E.B.S. i had to wait forever to see saved! too. mikalah gordon got kicked off american idol last night. while she wasn't the best singer, i wish …
Monday, March 28, 2005 - Miss Congeniality 2: Armed and Fabulous. I know this is overdue, but I wanted to use this opportunity to discuss an important topic. The issue at hand is known as the Sandra Bullock Effect (SBE). This theorem was first proposed by my brother, Arthur, so he is the real expert, but I will attempt to explain it here. The SBE is the degree to which any movie becomes watchable simply by the presence of a particular actor or actress who you happen to be fond of. For example, if I told you that someone made a movie about a clumsy, socially awkward, dowdy female police officer who goes undercover as a beauty pageant contestant to foil some impenetrable criminal conspiracy, you'd probably think to yourself, "Wow that sounds pretty dumb." And you'd be right. However…

Table 4: Typical references to movies in blogs: pre-release (top), and post-release (bottom).
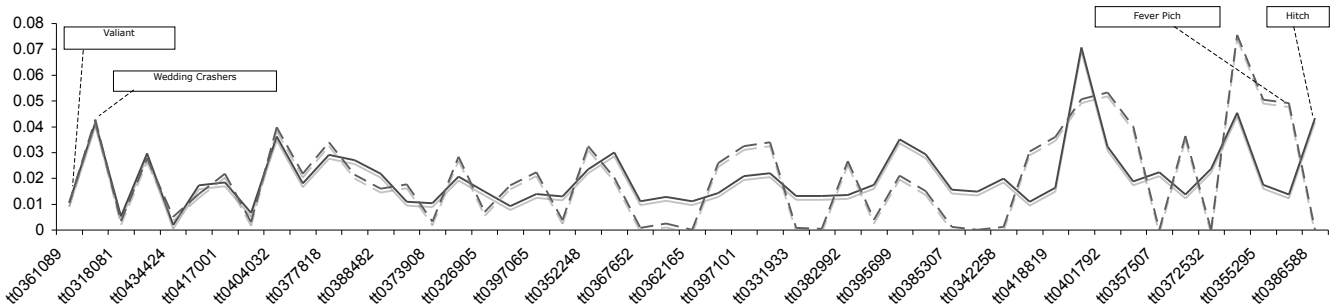

Figure 2: Per-movie comparison of income per screen (blue, continuous line) and positive references (green, dashed line), sorted by degree of correlation. For space reasons, the X-axis shows only the movie IMDB ID.

ponents of our analysis – the retrieval of posts from the collection of all posts, the polarity analysis, and so on – accumulates, and may destabilize the data.

Cursory examination of outliers in our experiments, both those that overestimate sales and those that underestimate them, did not yield any obvious underlying explanation.

Conclusions

We have shown that, in the domain of movies, there is good correlation between references to movies in weblog posts—both before and after their release—and the movies' financial success. Furthermore, we have demonstrated that shallow usage of sentiment analysis in weblogs can improve this correlation. Specifically, we found that the number of positive references correlates better than raw counts in the pre-release period.

In of itself, the correlation between pre-release sentiment and sales is not high enough to suggest building a predictive model for sales based on sentiment alone. However, our results show that sentiment might be effectively used in predictive models for sales in conjuction with additional factors such as movie genre and season.

References

K. Dave, S. Lawrence, & D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings WWW 2003*, 2003.

N. Glance, M. Hurst, & T. Tomkiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004*, 2004.

D. Gruhl, R. Guha, R. Kumar, J. Novak, & A. Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-135-X.

B. Liu, M. Hu, & J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings WWW 2005*, pages 342–351, 2005.

K. Nigam & M. Hurst. Towards a robust metric of opinion. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

K. Nigam & M. Hurst. Measuring aggregate opinion from text with confidence bounds. Intelliseek TR, 2005.

B. Pang, L. Lee, & S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings EMNLP 2002*, 2002.

J. Simonoff & I. Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13: 15–24, 2000.

R. M. Tong. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, 2001.

P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings ACL 2002*, 2002.