Box Office Forecasting Using Social Analysis

May 2013

www.infegy.com +1 816-494-1650 frontdesk@infegy.com
4151 N Mulberry Dr., Suite 240, Kansas City, Missouri 64116 USA

Have you seen "John Carter" or "Rise of the Guardians"? It's far more likely that you've seen "The Hunger Games," which nearly doubled its budget its opening weekend. Research shows peer influence is a strong motivator when it comes to movie choice; people trust their friends when they say a new flick is great, or when they advise to skip it. In the modern era, people are no longer limited to in-person small talk with close friends. Now, a huge amount of this interaction takes place online through blogs, forums, social media posts, and other sites, spreading via virtual word-of-mouth.

The potential popularity of a film is of massive concern to movie studios. Catching a likely flop before promotion ramps up, or capitalizing on an unexpected blockbuster before it hits theatres can translate into millions won or lost for each new release. For years, the film industry has relied on phone surveys and focus groups in attempts to forecast a film's success on its opening weekend. The limitations of those methods are well documented, but until recently there was no practical alternative. Now, in the age of social media, complete with the incredible explosion of open dialog online, is there a better, more accurate way of making these predictions? What if we could analyze that public conversation, interpreting commentary and opinion to gauge excitement for a new movie leading up to its premiere?

With access to Infegy Atlas, a powerful market research tool leveraging this immense wealth of online dialog, we can explore a more modern approach to forecast an upcoming movie's opening weekend. Based on our research, we hypothesize that through analysis of online consumer dialog in the seven days prior to a movie's release, we can model a more accurate prediction of the outcome of a box office opening.

To obtain the data for our study, we've taken advantage of the Infegy Atlas research platform, which is backed by a comprehensive database of online conversation. With a historical archive spanning over six years, we can look back into the conversations posted before each release, giving us a solid foundation on which to model our predictions. We can search the system to focus analysis around specific comments related to specific movies, with each query capable of analyzing millions of individual comments and opinions.

Because of the incredible volume of data posted to the web via blogs, social networks, forums and more, traditional manual analysis is impractical at best, and impossible at worst. The powerful analysis capability of Infegy Atlas tackles that challenge, automating the examination of the extensive amount of content available. The proprietary technology interprets written language, extracting a wealth of insights from natural dialog. And it does this fast, with the power to interpret tens of thousands of pieces of content in a single second.

Going deeper, this analysis moves beyond the simpler keyword analysis most common in the industry, harnessing cutting-edge linguistic analysis technology to gain true understanding of grammatical structure and content. Take, for example, the following snippet of text:

"I just saw Skyfall, and it was not bad! I would highly recommend it!"

For humans, understanding this content is easy. However, computer software must be taught, from the most basic level, how to interpret natural language. Infegy Atlas begins the process by breaking down this content to the the component parts. The subject, "Skyfall," is extracted and identified. Accurate subject-specific analysis is critical when analyzing the free-form text content found online. Even in cases where many subjects are mentioned, this process can identify all data relevant to the subject in question.

With the subject identified, the software begins its sentiment analysis. In the first sentence, many systems naively interpret the word "bad" as negative. With more advanced grammatical understanding, Infegy Atlas understands the negation from "not" modifies the phrase to be more positive overall. In the second sentence, "recommend" suggests a moderate positive sentimental opinion, but is then strongly elevated by the "highly" modifier. Overall, this pair of sentences would be interpreted as very favorable commentary for "Skyfall." For our purposes of validating our hypothesis, this robust sentiment analysis and volume measurement is more than sufficient.

With our source of consumer opinion ready, we need a set of movies to validate our hypothesis. Ten movies were chosen for testing, all opening during 2012. We used the top five openings of 2012: "The Avengers," "The Dark Knight Rises," "The Hunger Games," "Twilight: Breaking Dawn Part 2," and "Skyfall" to represent successful openings. The five unsuccessful openings are a bit more subjective: We looked for large-budget ($50 million or more) movies opening in 2012 in at least 2,000 United States theaters having less than 20 percent of their budget made on opening weekend. Even that list gives us more than five, so we looked for articles discussing "biggest losers" in box office openings over 2012, and found five most commonly appearing: "Rise of the Guardians," "John Carter," "Rock of Ages," "Total Recall" and "Dredd."

To perform our measurement, we gauge total positive conversation for each movie in the seven days before the movie opened. Infegy Atlas can gauge total volume within the "social universe", giving us a projection of the volume of all online dialog during a period over an abundance of channels. However, volume of conversation alone isn't enough to gauge popularity. If there is a large volume of conversation, but the majority of it is negative towards the movie, we believe this would imply a poor opening. So, we also deeply interpret that volume of dialog, and extract two additional key figures. First, we look at how often commentary about a film is sentimental, or expressing an opinion. Second, within those we can see what percentage of comments are favorable, or positive, using the process described above.

Now we have identified three key statistics based on consumer dialog for each film. Taking our original total volume number, scaling by the sentimental percentage, and again by sentimental positivity, we have a projection based on the total positive comments for each movie in the seven days leading up to its release. This gives us a solid metric with which to test our hypothesis.

Each of these numbers by themselves do not offer much insight into box office performance. As with most analysis, context is crucial. For our examination here, we'll look at how these numbers compare both with each other, and how the social data correlates with the box office openings. We'll look at the latter by measuring total opening weekend revenue divided by total positive social volume, giving us a dollars-per-comment ratio which we can compare movie-to-movie for measurement of consistency of this relationship.

Our findings:

	Opening ($/millions)	**Opening Date**	**Positive Commentary**	**$/PC**
The Avengers	$207.4	May 4	864,102	$240.02
The Dark Knight Rises	$160.8	July 20	669,460	$240.19
The Hunger Games	$152.5	March 23	561,286	$271.70
Twilight: Breaking Dawn II	$141.1	November 16	375,194	$376.07
Skyfall	$88.3	November 9	261,836	$337.23
Rise of the Guardians	$32.6	November 21	137,735	$236.69
John Carter	$30.2	March 9	115,748	$260.91
Rock of Ages	$14.4	June 15	44,984	$320.12
Total Recall (2012)	$9.2	August 3	34,840	$264.06
Dredd	$6.3	September 21	22,155	$284.36
			Average:	$283.13
			Deviation:	$46.88

From this chart, it's immediately apparent the ordering for the box office opening matches the ordering for total positive commentary. Looking deeper, we see that our ratio of opening dollars per positive comment remains fairly consistent, averaging $283 and deviating by just $47. This is fairly remarkable, considering the variety of movies analyzed here. Take "John Carter," for example. This movie was backed by a massive $250 million production budget, with its $100

million marketing budget alone surpassing all total costs for most major releases. However, all that money wasn't enough to make it a hit, with the film's relatively paltry $30 million opening weekend earnings.

In contrast to "John Carter," "The Hunger Games" released after a much smaller $78 million production budget and $45 million marketing budget. Despite having a third the money backing it, however, "The Hunger Games" had an opening weekend bringing in five times the ticket sales!

Through this example, marketing budgets certainly aren't an accurate measure of forecasting movie popularity, but how'd our social data fair? Examining the positive commentary, "The Hunger Games" racked up around 561 thousand positive mentions in the week leading up to release, while "John Carter" didn't do quite so well with about 115 thousand. The ratio here is 4.8-to-1, nearly identical to the difference in openings at 5.0-to-1.

This pattern holds up throughout the other eight films, with our social data reliably forecasting all 10 openings with simple linear math on positive commentary from the preceding week. This seems to validate the hypothesis we set out to test: Social commentary shows strong potential for accurately forecasting box office openings.

Can this data be applied to forecasting other things as well? Based on our experience, it's a resounding yes. At Infegy, we've used this data for forecasting for financial markets, political elections, even accurately forecasting the winner of American Idol weeks in advance, all in similar fashion to the process described here.

In 2013, the wisdom of the crowds is more powerful than ever. With online conversation growing every day, it is an incredible resource for predictions, opinions, suggestions, advice, and commentary. Harnessing this incredible resource has proved to be beneficial to help decide which political candidate best represents our beliefs, which stocks to invest in, or what movie will be the next blockbuster. The opportunities are limitless and too powerful to ignore.