Business Intelligence from Social Media: A Study from the VAST Box Office Challenge

Yafeng Lu, Feng Wang, and Ross Maciejewski, *Member, IEEE*

Abstract—With over 16 million Tweets per hour, 600 new blogs posts per minute and 400 million active users on Facebook, businesses have begun searching for ways to turn real-time consumer based posts into actionable intelligence. The goal is to extract information from this noisy, unstructured data and use it for trend analysis and prediction. Current practices support the notion visual analytics can play a large role in enabling the effective analysis of such data. However, empirical evidence demonstrating the effectiveness of a visual analytics solution is still lacking. This paper presents a visual analytics system which extracts data from Bitly and Twitter to use for box office revenue and user rating predictions. Results from the VAST Box Office Challenge 2013 demonstrate the benefit of an interactive environment for predictive analysis compared to a purely statistical modeling approach. These visual analysis method used in our system can be generalized to other domain where social media data is involved, such as sales forecasting, advertisement analysis, etc.

Index Terms—social media, box office, visualization, prediction

1 Introduction

Social media data presents a promising, albeit challenging, source of data for business intelligence. Customers voluntarily discuss products and companies, giving a real-time pulse of brand sentiment and adoption. Unfortunately, such data is noisy and unstructured, making it difficult to easily extract real-time intelligence. Thus, the use of such data can be time-consuming and cost prohibitive for businesses. One promising current direction is the application of visual analytics. Recently, the visual analytics community has begun focusing on the extraction of knowledge from unstructured social media data [12]. Studies have ranged from geo-temporal anomaly detection [3], [4] to topic extraction [14] to customer sentiment analysis [5]. The development of such tools now enables end-users to explore this rich source of information and mine it for business intelligence.

One key area for business intelligence is revenue prediction. One means of revenue prediction is utilizing social media to understand product adoption and sentiment. Currently, very few tools exist that effectively enable the exploration of social media (such as Twitter) in conjunction with traditional business intelligence analytics (such as linear regression). Due to the abundance of social media discussions on movies, movie revenue prediction has drawn much attention from both the movie industry and academic field. Movie meta-data, social media data and google search volumes have all been explored in various prediction methods. For example, an early study by Simonoff et al. [13] predicted box office revenue with a logged response regression model using meta data features (e.g., time of year , genre, MPAA rating) as categorical regressors. Zhang et al., [15] demonstrated that regression models based on meta data features can be enhanced by utilizing variables extracted from news sources, and Joshi et al. [6] explored the relationship between film critic reviews and box office performance. Further work by Asur et al. [1] found that the rate of Tweets per day could explain nearly 80% of the variance in movie revenue prediction, and recent work from Google [10] claimed a 94% prediction accuracy in box office prediction by utilizing the volume of internet trailer searches for a given movie title.

While such methods have demonstrated the benefits of social media for extracting business intelligence for box office revenue prediction, they have relied solely on automated extraction and knowledge prediction. This paper presents our visual analytics toolkit for movie box office prediction. Our toolkit consists of a web-deployable series of linked visualization views that combine statistical techniques (multiple linear regression and time series modeling) with data mining (sentiment analysis) for predicting the opening weekend gross and viewer rating scores of upcoming movies. This type of visual analytics approach for social media analysis and forecasting can be directly applied to a wide range of business intelligence problems. Understanding how information is spread as well as the underlying sentiment of the messages being spread can give analysts critical insight into the general "pulse" of their brand or product. Developing a set of quick look visualization tools for an overview of such social media data along and linking this to models that business analysts generate for deploying new products, advertising campaigns and

- *Yafeng Lu, Feng Wang, and Ross Maciejewski, are with the School of Computing, Informatics and Decision Systems Engineering at Arizona State Unviersity.*
  E-mail: {lyafeng, fwang49, rmacieje}@asu.edu.
- *Visual Analytics and Data Exploration Research (VADER) Lab - http://vader.lab.asu.edu*

sales forecasts can be critical. Our toolkit can also be used to explore other business related social media data, for example, to see how well an ads campaign did and the pattern of information spreading. Some exploration can help adjust business decisions.

In order to demonstrate the effectiveness of our system, this paper reports on the results of the Visual Analytics Science and Technology (VAST) Box Office Challenge 2013. Results from this challenge also allowed us to explore the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution. Our results demonstrate that our analytics team was able to outperform the purely statistical model solution during the course of this contest; however, results from this study merely support the hypothesis that visual analytics can improve an end-user's analytic capabilities. More studies are required to create further convincing evidence.

2 Data Extraction, Analysis and Visualization Tools for Box Office Predictions

In order to explore the impact that visual analytics can have on generating insight into social media data, our work focused on box-office predictions using Twitter indices, bitly links, and access to the Internet Movie Database. This system is a web-enabled visual analytics toolkit that allows analysts to quickly extract, visualize and clean information from social media sources. These tools were combined with linear regression and temporal modeling for movie box office prediction and sentiment analysis for movie review rating prediction. In this section, we will discuss the various tools developed as well as lessons learned from the contest.

2.1 Tweet Mining – Overview, Sentiment and Cleaning Tools

While the tools developed are applicable to a variety of social media analysis problems, our specific application focused on structured data from the internet movie database (e.g., genre, budget, rating), and unstructured data from social media (e.g., Tweets, blog posts). While structured data is relatively straightforward to extract, unstructured data requires a large amount of pre-processing and manipulation. Unstructured data collected from social media revolved around movie related Tweets and bitly URLS. Tweets were collected for the two-week period prior to the release date based off the hashtag provided by a movie's official Twitter account. Our goal was to develop tools that could extract a variety of metrics from Twitter and IMDB (see the summary in Table 1 of the metrics we found most useful). Several of the extracted metrics required data mining and cleaning. To facilitate this, we developed tools that could present the volume of Tweets at various levels of temporal aggregation(Figure 1 (a)), enable users to remove unrelated Tweets from the aggregate metrics, and allow users to extract and manually adjust the sentiment of a Tweet (Figure 1 (b-d)).

TABLE 1: Variables Description

Variable	Description
OW	3-day Opening Weekend Gross
Budget	Approximate movie budget from IMDB. (unit is "million" of dollars)
Genre(category)	The movie's genre(s) according to IMDB
TUser	Number of unique users who tweeted about a movie
TBD	The average daily number of Tweets over the 2 weeks prior to release
TSS	Tweet Sentiment Score - A summation of each individual word's sentiment polarity as calculated via SentiWordNet [2]
MSS	Movie Sentiment Score - A derivation of the overall sentiment of a movie
MSP	Movie Star Power - A summation of the Twitter followers of the three highest billed movie stars (as listed by IMDB)

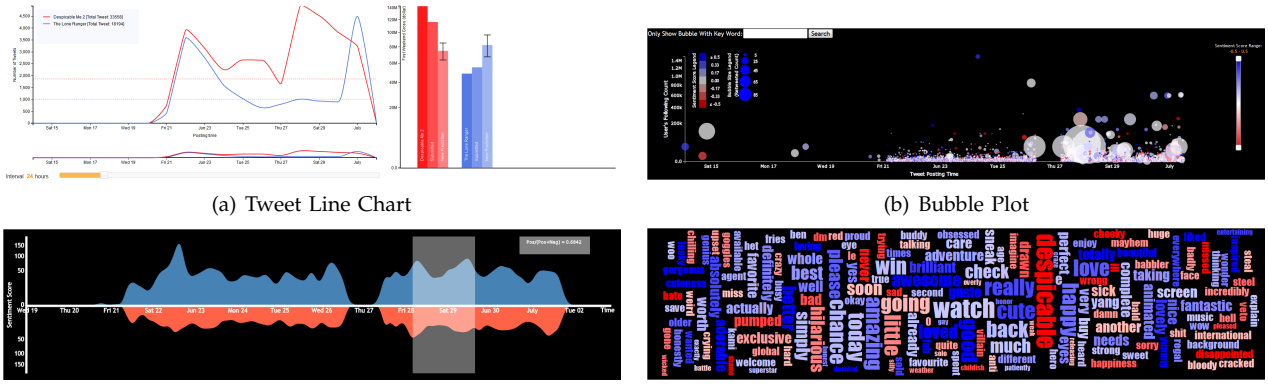
In order to approximate the popular sentiment of a movie, we processed each Tweet using a dictionary based classifier, SentiWordNet [2]. This process assigns each word in the Tweet with a score from -1 to 1 with -1 being the highest negative sentiment score and 1 being the highest positive sentiment score. Next, each Tweet is assigned a sentiment score by summing the sentiment score of all words in the Tweet and scaling the range from $-.5$ to $.5$ (TSS in Table 1). Finally, the movie sentiment score (MSS in Table 1) is calculated as

$$MSS = \frac{Positive\ Score}{Positive\ Score + Negative\ Score} \quad (1)$$

where *Positive Score* is the sum of all Tweets for a given movie with a TSS greater than zero and *Negative Score* is the absolute value of the sum of all Tweets for a given movie with a TSS less than zero.

Once the sentiment scores for Tweets were extracted, these values were then visualized to the end user. Figure 1 (b-d) shows the bubble plot view, the sentiment river view, and the sentiment wordle view. In the sentiment wordle view (Figure 1 (d)), the 200 most frequently mentioned words are extracted and visualized.

Both the bubble plot and the wordle plot enabled interactive searching and filtering by keywords and users. Users posting irrelevant messages could be removed from the Tweet count and mismatched sentiment could be modified by the end user. The primary use we found for the views in Figure 1 were for data cleaning. The primary lesson learned was that visualization tools are a necessity for data cleaning due to the noisiness of social media data and the problems inherent in sentiment matching using a sentiment dictionary (e.g., phrases such as "I want to see this movie so bad" are marked as negative due to the word "bad", and words such as "Despicable" give negative sentiment when they are merely references to a movie title). While the wordle view provided a quick way to assess the sentiment of popular words, it was necessary to hover over the bubble



(a) Tweet Line Chart

(b) Bubble Plot

(c) Sentiment River

(d) Sentiment Wordle

Fig. 1: Tweet trend and sentiment views for Despicable Me 2. (a) Line charts and bar graphs showing the number of Tweets per day and the predictions. (b) A Tweet bubble plot where blue represents positive sentiment and red represents negative. The size of the bubble represents the number of times a Tweet has been retweeted, the x-axis is time, and the y-axis is the number of followers that the user who submitted the Tweet has. (c) A sentiment river view where sentiment is aggregated over four hour intervals. Positive sentiment is plotted in red above the x-axis, negative in blue below. A user can select an area on the river to see the ratio of positive to negative sentiment. (d)A sentiment wordle where the size of the word represents the number of times it was used in a Tweet and color represents sentiment. By clicking a word, the bubble chart view will be filtered to only Tweets containing that word.

plot or open a Tweet list view through the search bar in order to fully explore the context of a Tweet. While such views were useful for data cleaning, our analysis approach (see Section 3) demonstrated to us that these views were more effective for cleaning and overview than for use in the model analysis. The critical need for tools to extract the correct metrics for regression modeling is a major hurdle that needs to be overcome in utilizing social media data for business intelligence. The bubble plot and wordle plot helped us to deal with the challenge of sentiment analysis and cleaning of noise from social media data.

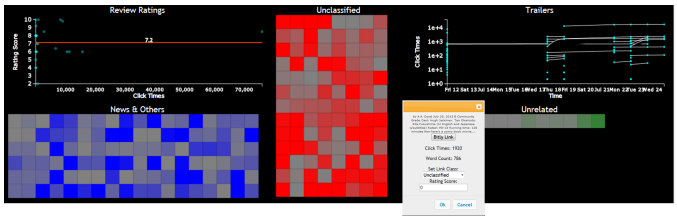


Fig. 2: Our interactive bitly classification widget. In the center are the unclassified links which the user can click and classify as seen in the floating window. The upper left is a plot of review score by click counts with a line for the average review score value.

2.2 Bitly Mining

While Tweets could be reasonably processed via the SentiWordNet dictionary, blog posts required a different approach. As part of this work, we explored long-form text by extracting bitly links containing movie keywords. These links typically consisted of review articles or news reports about the movies (or in many cases unrelated news, for example when the movie "The Heat" was released, the Miami basketball team, The Heat, had just won the NBA championship). For our review score prediction, we relied on prescreening review scores that were embedded in bitly links and developed an interactive tool for extracting these scores as shown in Figure 2. Initially, each bitly link starts as unclassified and is represented in a pixel matrix (color saturation corresponds to the number of times a link was clicked). By clicking on an unclassified square, a pop-up box appears with a brief bit of text from the article. The user can then choose to follow the link to scan the article for review scores and then manually assign a review score to an article or classify it as news or unrelated. A plot of review scores from articles versus the number of times an article was accessed is provided for analysis (see the upper left quadrant of Figure 2). This tool allows for quick data filtering and extraction, for example, reviews for the Star Trek video game can easily be separated from the Star Trek movie which would be difficult to automatically encode. Furthermore, the color coding from the pixel matrix can be used as a metric for classifying only those articles that had a substantial amount of views.

Similar to our lessons learned in Tweet mining, extracting information from bitly can be difficult to fully automate. As in the Star Trek example, multiple products for a movie may be released and reviewed at the same time. Furthermore, review scores may range from "two thumbs up" to "4 out of 5 stars" to "6 out of 10". With the analyst in the loop, these scores can be mapped to a user's own base system (in this case our metric was out

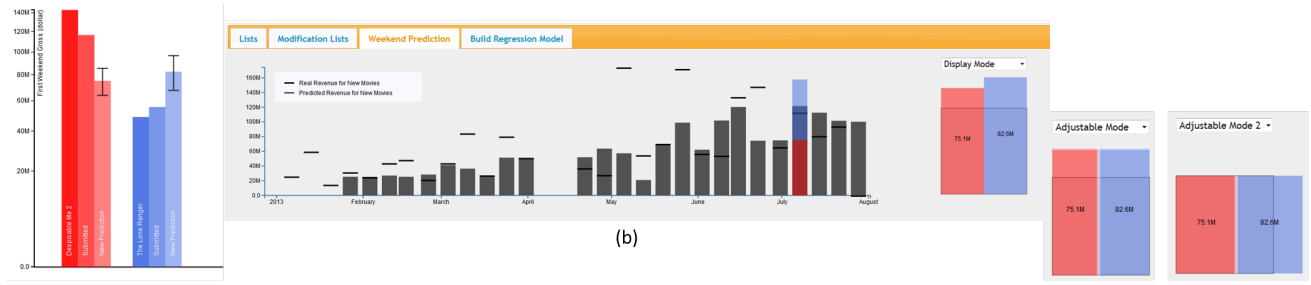


Fig. 3: The weekend prediction view for newly released movies and the prediction adjustment widget. This is the weekend when Despicable Me 2 and The Lone Ranger were released. (a) The bar graph view showing the actual value, submitted prediction and model prediction. (b) The stacked graph view showing the predicted weekend gross overlaid with the upcoming movie's regression model prediction. (c) The adjustment widget where users can modify the gross prediction; however, the predicted values for the new movies remain proportional. (d) The adjustment widget for changing individual predictions. The gray box represents the total weekend gross.

of 10).

2.3 Regression Modeling

Once data cleaning and variable extraction was complete, the next task was to use the social media metrics to develop a model for predicting box office revenue and review scores. Traditional variables used in these box office prediction models include structured variables (e.g., MPAA rating,movie budget) and derived measures (e.g., popularity of the movie stars, popular sentiment regarding the movie). Based on our initial literature search, we chose to utilize multiple linear regression for an initial prediction range for the opening weekend box office revenue (see the sidebar for a brief introduction to multiple linear regression modeling). We explored a variety of different variables that could be mined from the contest, see Table 1. After initial model fitting and evaluation using R [9], we found our best fit to be of the form:

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget + \varepsilon \qquad (2)$$

The model is updated weekly as new movies are entered into the data set. Parameters are fit using movie data beginning in January, 2013. Our first prediction was for the May 17th weekend and used data from 39 movies for training. Our weekly models reported an $R\text{-}adj^2 \approx 0.60$ with $p < .05$. Our final parameters were $\beta_0 \approx 4.9 \times 10^3$, $\beta_1 \approx 4462$, and $\beta_2 \approx 2.3 \times 10^5$.

The drawback of this model is that it does not fit the data overly well and predictions have a large variance. For comparison, a linear regression model using google search volumes was reported to explain more than 90% of the variance on box office performance [10], and models by Asur et al. [1]also report an $R\text{-}adj^2$ of over 90% when the number of theaters was used as a regressor. Our hypothesis was that a visual analytics toolkit could partially enable analysts to overcome poor data (partially due to the noise in social media data and partially due to the closed world nature of the contest). In order to facilitate better model prediction, we created a simple bar graph view (Figure 3(a)) which, for historical movies, showed the model prediction and its 95% confidence interval error range, our submitted prediction, and the actual box office gross. For new movies, only the model prediction and user submission was shown. This view was critical in our analysis process, and the primary view into the data consists of an overview of the Tweets per day and the model predictions of the movies under analysis as shown in Figure 1(a).

2.4 Temporal Modeling

While the regression model is able to provide one point for analysis, our goal was to also provide a big picture overview. For any given weekend, there is likely a maximum amount of money available in the market. In order to approximate the total amount of money available in the market, we employed a simple moving average model. Limitations here included access to data (historical weekend grosses were not available, and after a movie opens, further weekend takes were no longer reported in the contest). To compensate for this, we approximated subsequent weekend grosses for movies under the assumption that movies would run for three weeks following their opening weekend, and each weekend their box office take would be reduced by 50%. Thus, for any given weekend, we approximated the gross as:

$$\textit{Weekend Gross}(t) = \sum_{\forall_i} \mathrm{OW}_i(t) + \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t-j),$$

where t is the current weekend and i is the index to a movie that exists at time t. Then, for the weekend gross prediction, we use a moving average:

$$\textit{Weekend Gross}(t+1) = \frac{1}{3} \sum_{j=0}^{j=2} \textit{Weekend Gross}(t-j).$$

Finally, we approximate the available revenue for new movies as:

$$\textit{New Movie Gross}(t+1) = \textit{Weekend Gross}(t+1) - \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t+1-j).$$

While this prediction is crude, it provided the analysts with a valuable bound in which to explore the revenue predictions.

Results from the temporal weekend prediction and the linear regression models were then visualized in two different views as shown in Figure 3. The first view consists of a linked bar graph combined with stacked bars as shown in Figure 3 (b). The primary portion of the bar graph consists of light gray bars indicating the predicted total weekend market for the new movies and the dark gray short line indicates the actual weekend market for each calendar week whose date is shown on the x-axis. The stacked color bar graph is visualized only for the weekend under analysis, and the color design is the same as the movie's color in the prediction bar graph.

The second view, Figure 3 (c) and (d), is used to enable users to interactively adjust predictions while also visualizing the bounds of the total weekend prediction. In this view, a gray square is drawn, the area of which is scaled linearly to the total weekend prediction. Colored rectangles are superimposed onto the gray square, where the area of each colored rectangle represents the linear regression prediction for each movie being released on that weekend. If the sum of the individual predictions is equal to the total prediction, the colored rectangles will fit exactly into the gray square in both Figure 3 (c) and Figure 3 (d). The color design is the same as those of the bar graph, and modifying the size of a bar in any view will modify the size across all views.

Our system was designed to allow for three types of prediction adjustments.

1) Users are allowed to change the amount of the total gross prediction but the ratio between the movies will remain consistent.
2) Users are allowed to change the amount of an individual prediction but the total weekend prediction is kept consistent.
3) Users are allowed to arbitrarily change each movie's prediction and ignore the weekend gross.

By implementing and integrating multiple comparison methods, we found that we were able to quickly bound our analysis. While flexible, these bounds provided us with an early estimate of the total expected weekend gross in which to compare the predictions of our linear regression models. This multiple model comparison was a critical step for our overall box office prediction and was regularly used for all movie analyses.

While the results of our temporal predictions were of low quality, the combination of predictions and bounding of the problem space provided critical information for comparison and analysis. We will further discuss in Section 3 how the combination of both models was critical for successful predictions. Overall, the addition of multiple models predicting similar information can help guide analysts to a better ground truth. Similar to principles employed in the delphi method [11], where predictions are solicited from multiple experts and used to come to a common conclusion, in our system, we allow users to solicit predictions from multiple models to aid in their analysis. This bounded adjustment widget can be used in other hierarchical predictions which have both individual and total predictions, such as sub-topic trend prediction in a time period.

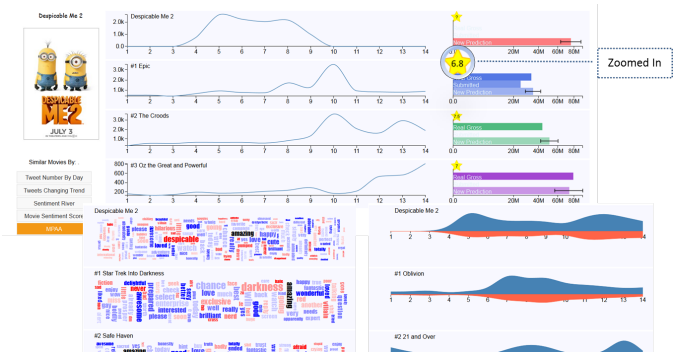


Fig. 4: A user defined similarity view cropped to show the topmost similar movies. In the center is the Tweets by day view, on the right is a graph of the opening weekend gross. There are bars for the actual gross, our final prediction, and the prediction range. The star in the upper left corner of the graph shows the review score.

2.5 Similarity Visualization

While bounding the movie predictions provided context for an overview of the total weekend, our other critical analytic view was the similarity widget. This widget enables analysts to quickly find and compare the accuracy of prediction based on various criteria of similarity. This allows analysts to determine if the given prediction model typically underestimates, overestimates or is relatively accurate with regards to movies that the analyst deems to be similar. In this manner, a user can further refine their final prediction value for both the box office gross and the review score. In this work, we have defined nine similarity criteria with distance calculation methods defined in Table 2. In all similarity matches, we show the top five most similar movies. These views allow users to directly compare Tweet trends and sentiment words between movies deemed to be similar in a category. Figure 4 contains snapshots from the Despicable Me 2 similarity page showing the line chart view with an MPAA similarity criterion, a wordle view with top word similarity criteria and a theme river view with sentiment similarity criteria.

While all of the variables used in our similarity metric could also be used in the linear regression model, results of the modeling indicated that these variables were not significant in altering the model. However, by providing an analyst with insight into these secondary variables,

TABLE 2: Calculations of Similarity Criteria

Similarity Criteria	Distance Measurement
Tweet Number by Day	$Dis(v,s) = \sum_{i=1}^{14} \lvert TBD_i(v) - TBD_i(s) \rvert$
Tweet Changing Trend	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{TBD\_i(v)}{Max(TBD\_j(v), j=1,2,...14)} - \frac{TBD\_i(s)}{Max(TBD\_j(s), j=1,2,...14)} \right\rvert$
Sentiment River	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{MSS\_i(v)}{Max(MSS\_j(v), j=1,2,...14)} - \frac{MSS\_i(s)}{Max(MSS\_j(s), j=1,2,...14)} \right\rvert$
MSS	$Dis(v,s) = \lvert MSS(v) - MSS(s) \rvert$
MPAA	same MPAA rating and close release date
Genre	$Dis(v,s) = 1 - \frac{card(Genre(v) \cap Genre(s)) \times 2}{card(Genre(v)) + card(Genre(s))}$
MSP	$Dis(v,s) = \lvert MSP(v) - MSP(s) \rvert$
Sentiment Wordle	$Dis(v,s) = 1 - \frac{card(SWordle(v) \cap SWordle(s))}{card(SWordle(v))}$

coupled with the temporal weekend modeling, further refinement of the prediction is made possible. For example, an analyst may compare the absolute difference between Tweets of two movies, or they can inspect the trend of the Tweets through line chart comparison using the Tweets Changing Trend similarity metric. This tool also allows users to quickly compare the current movies under analysis to recently released movies with the same Motion Picture Association of America rating, genre or movie stars popularity based on the number of Twitter followers a star has.

3 A Visual Analytics Process for Box Office Predictions

This system was used to predict 23 movies over the course of 3 months in the VAST 2013 Box Office Challenge. Our prediction process involved 3 steps. Our example prediction process focuses on the July 4th holiday in the United States when Despicable Me 2 and The Lone Ranger were released.

3.1 Movie Review Score Prediction Process

Our movie review score process centered around using the wisdom of the crowd for predicting an expected IMDB review score. For each movie, our process began by entering the bitly view and manually extracting review scores from bitly users who had done a prescreening of the movie (Figure 2). In the case of Despicable Me 2, the analysts manually classified the most clicked bitly reviews. The average value of all review scores extracted for Despicable Me 2 was 7.8. Once the average value is recorded we would then use the similarity view to compare to other movies. The movie review score is visualized as a star highlighting the review value in the corner of the bar graphs (Figure 4). Typically we would compare across genre, movie rating and sentiment to determine if we felt the average value extracted from bitly links was a reasonable prediction. In the case of Despicable Me 2, we compared to Monsters University as both movies were animated sequels. Monsters Universitys IMDB rating was 7.8 giving us confidence that our predicted value of 7.8 was reasonable. This same process was then performed for the Lone Ranger, and a viewer rating of 6.4 was predicted.

3.2 Movie Gross Prediction Process

Once the viewer rating was predicted, we then focused on determining the box office gross for the two movies. This weekend was challenging for two reasons. First, the data stream from the contest was broken, providing only 6 days worth of Tweets, and, second, the predictions were for a five-day weekend as opposed to the typical three-day weekend. Using the available data, we obtained a rough estimate for the Despicable Me 2 box office value in the range of $76M +/- $13M and $85M +/- $13M for The Lone Ranger. Next, we explore the expected three-day weekend total and see that our time series model approximates that $124M is available for the two movies for the three-day weekend. A quick look at Figure 3 shows that our regression predictions are well outside the bounds of the time series model prediction.

Given the misalignment between the two models, we begin exploring the similarity views to determine which movies The Lone Ranger and Despicable Me 2 are most similar to based on our predicted review score as well as various other metrics. We compare Despicable Me 2 to a variety of animated movies and we see that the predicted $73M is actually low when compared to animated movies such as Monsters University. Next, we explore various similarity views for The Lone Ranger and see that it is likely similar to World War Z, which had a weekend gross of $66M. However, World War Z's viewer rating was much higher at 7.4 than the predicted 6.4 for The Lone Ranger.

After looking at the available information, we determined that Despicable Me 2 should perform similarly to Monsters University, and we predicted a three-day gross of $85M. Based on our temporal prediction, this left only $39M for The Lone Ranger; however, given the other evidence, it seemed likely that The Lone Ranger would underperform. Finally, we took our three-day prediction values and linearly scaled them to be a five day prediction, resulting in a final five day prediction of $116.5M for Despicable Me 2 and $55.45M for The Lone Ranger. The actual three-day gross for Despicable Me 2 was $83.5M and $29M for The Lone Ranger. The actual five-day gross for Despicable Me 2 was $143M and $48.7M for The Lone Ranger, and the actual IMDB ratings were 7.9 for Despicable Me 2 and 6.8 for The Lone Ranger.

TABLE 3: Comparison with Peer Teams Predictions

Team	Gross Prediction				Viewer Rating			
	Entry	Average Error	STD	MRAE	Entry	Average Error	STD	MRAE
VADER(Interactive)	23	11.213	9.416	0.467	23	0.487	0.460	0.075
Team Prolix	23	16.466	15.195	0.424	20	0.82	0.640	0.129
Uni Konstanz Boxoffice	14	17.056	15.743	3.929	21	0.905	1.519	0.095
CinemAviz	21	17.219	17.677	1.970	21	0.738	0.559	0.114
Team Turboknopf	8	21.9	15.606	0.685	18	0.514	0.426	0.079
elvertoncf - UFMG	3	12.677	9.806	3.009	3	1.323	0.328	0.259
Philipp Omentisch	5	30.657	38.028	0.678	5	0.5	0.324	0.071
CDE IIIT	2	60.6	62.084	0.537	2	0	0	0

4 Results From VAST Challenge

Eight teams (our team being Team VADER) from various research institutes participated in the VAST Box Office Challenge. Data was also collected from 4 professional movie prediction websites. In this section, we compare our prediction performance with respect to peer teams from the VAST challenge and professional predictions.

4.1 Comparison with Peer Teams

Table 3 provides summary statistics of the performance of each team that participated in the VAST Box Office Challenge. For the gross prediction we report the average error (in terms of millions of dollars), the standard deviation (STD) of the average error term and the mean relative absolute error (MRAE), which is the percentage of bias deviating from the real value.

$$MRAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Prediction_i - RealValue_i|}{RealValue_i} \quad (3)$$

Similar values are reported with regards to predicting the IMDB rating (in the case of the IMDB rating, participants submitted a rating score from 1-10). These statistics can be interpreted by their magnitude, where smaller values indicate more accurate predictions. Data collected in Table 3 was provided to all challenge participants after the contest was closed.

In terms of average error and standard deviation, our team reported the lowest values in gross prediction across all teams. With respect to the MRAE for gross prediction and viewer rating, our results are slightly worse than Team Prolix (MRAE of .424 for Prolix compared with our .467), and similar in range to Philipp Omentisch, CDE IIIT and Team Turboknopf. While Team Prolix was able to achieve a smaller MRAE over the contest than our group, comparatively, they have a much larger average error and standard deviation indicating more inconsistency in their predictions.

With regards to the viewer rating prediction, our team had the lowest average error and MRAE of all teams with more than 5 submissions. CDE IIIT submitted two perfect predictions; however, those were CDE IIIT's only predictions making it difficult to determine if their methods would produce consistent results. With regards to the average error and standard deviation of the viewer rating, our team had similar results to Team Turboknopf, slightly besting them with regards to Average Error, but being slightly worse with regards to standard deviation.

TABLE 4: Comparison with Professional Predictions.

Prediction Source	Entry	Average Error	STD	Average MRAE
VADER (interactive)	21	12.729	9.425	0.285
VADER (No interaction)	21	23.051	22.011	0.501
boxoffice.com	21	8.538	7.466	0.191
filmgo.net	6	12.75	7.409	0.297
hsx	20	9.06	7.397	0.205
boxofficemojo	14	9.864	7.527	0.224

4.2 Comparison with Professional Predictions

In order to explore the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions we have also collected results from four professional prediction websites for comparison. For our comparison to the professional prediction websites, we again explore the results of the VAST Box Office challenge. Given that these results were collected and verified by the contest organizers, we feel this is an adequate means of justifying their validity. For the comparison in Table 4, only 21 movies are shown in the chart as two movies, The Bling Ring and The To Do List, were limited release movies which opened in only 5 and 591 theaters respectively and most expert prediction sites do not provide predictions for limited release movies. For each prediction, we followed the same general process as described in Section 3. As previously stated, the underlying linear regression model used in our system was significant with an R^2-$adj \approx .6$.

Results in terms of the MRAE are given in Figures 5 and 6 for the opening weekend gross and review score respectively. Figure 5 provides a comparison of our MRAE with that of several expert prediction websites. From Figure 5, it is clear that we outperformed the experts in the case of three movies (Epic, Hangover 3 and Fast and Furious 6), and in the case where we had the largest error (After Earth) we relied heavily on the analytical component with no interaction.

Table 4 gives the average error, standard deviation and MRAE for the predicted movies. What the results show is that for the model used, the predictions of our team utilizing an interactive tool were a dramatic

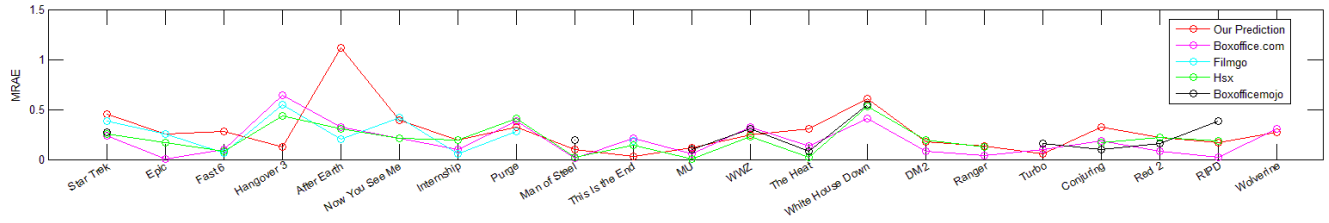


Fig. 5: The mean relative absolute error of box office weekend gross predictions, where the x-axis is the predicted movies.

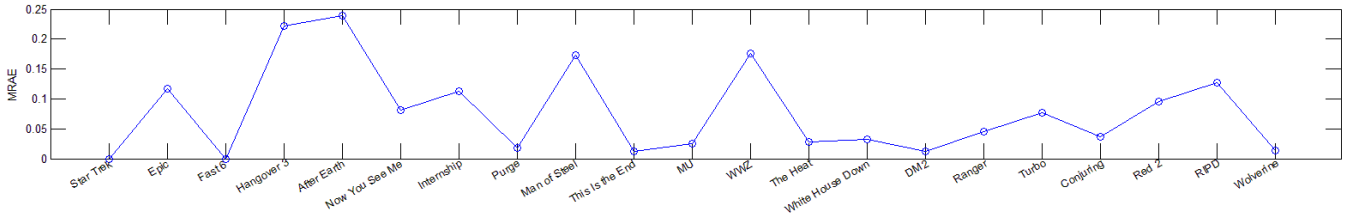


Fig. 6: The mean relative absolute error of our viewer rating predictions, where the x-axis is the predicted movies.

improvement over just the model itself (see Table 4 VADER (Interactive) versus VADER (No Interaction)). This provides a strong indication that the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution is valid. However, we do not wish to overstate our claims. This contest provides only a single data point for exploring how one group of analysts in a closed world setting were able to utilize a visual analytics toolkit for improved prediction. What this demonstrates is the need for further controlled studies in which a group of analysts perform similar model predictions and results are compared between analysts using a visual analytics platform and analysts using only results from a given regression model. However, results from the contest indicate that a visual analytics toolkit can enhance business intelligence.

Further analysis of the data also indicates that these tools enabled our team of novice box office analysts to quickly close the gap between the experts. Table 4 shows the average error and standard deviation for our predictions and compares them to four well known professional prediction websites. What we see is that both our average error and average MRAE are slightly lower than filmgo.net indicating that our methodology enabled our group of novice analysts to be competitive when compared to expert analysts. The significance of this relies on three major assumptions:

1) The professional prediction websites have more experience in box office prediction than our team.
2) The professional prediction websites have access to more data than our team was allowed in the closed world contest.
3) Access to more data can enable better predictive models as evidenced by [1], [6], [10], [15]

First, it seems reasonable that a professional prediction website would have much more experience than a computer science team who has never previously attempted to predict box office sales. Second, it is clear that utilizing data sources (specifically the number of theaters a movie is released in) will result in a better prediction model (a larger R^2). From these assumptions, it becomes clear that (in this instance) the application of a visual analytics toolkit can enable individuals that are knowledgeable with respect to data analysis to quickly understand information being presented to them in new domains and make predictions that are in line with expert predictions. Overall, our prediction error (.285) was slightly lower than that of filmgo (.297), but approximately 50% worse than boxoffice.com (.191). However, if we remove the After Earth and Now You See Me weekend (during which we relied heavily on the model and very little on the interactive visuals), our MRAE drops to .239 which puts us near the prediction range of boxofficemojo. Other sources of error can be accounted for in disrupted Twitter and bitly data feeds. These interruptions were pronounced for The Heat, White House Down, Monsters University and World War Z. However, even with those interruptions, our predictive analysis process was still quite robust with only The Heat being a significantly worse prediction than the professional sites.

5 Conclusions and Future Work

Overall, the application of visual analytics for social media analysis has proven relatively effective. However, there are still many challenges in applying this to all domains of business intelligence. First, social media data is extremely noisy. Movie predictions work well as one can track the effectiveness of ad campaigns by following the specific hashtags promoted by a brand. As

the analysis gets farther afield from Twitter (for example when trying to mine data from bitly) it becomes difficult to choose effective keywords. Second, due to the ever changing stream of social media sources and users, it is likely that any automated system for data collection and prediction will eventually be steered off course. As such, it is critical to link the human into the loop; however, as is evidenced by the issues in sentiment analysis, the data cleaning process should not overburden the analyst. The sentiment analysis and cleaning process employed in this work places an overly large burden on the end user. As such, integrating a system for having a user label a subset of tweets for sentiment model training could be a more effective solution. Third, it is imperative to link highly curated small datasets with this so call "big data". While social media data can be used as a proxy for many signals, we find that linking multiple data sources with varying levels of reliability (for example, total weekend take for all movies and regression modeling) can enhance the predictive abilities of a system. For example, doing focus groups and linking their data with results from social media could enhance the analysis of a proposed new product release. Finally, this paper demonstrates the need for interactive tools to mine social media data. From the examples of box office prediction, it is clear that such data contains a wealth of information. However, extracting knowledge from this data and effectively communicating this remains a challenge. There are clear needs for effective data cleaning tools to improve filtering of unrelated social media signals, as well as for improving the results of challenging analytical problems (such as sentiment analysis). Our results demonstrate that the use of visual analytics tools can have a significant impact on knowledge discovery for business intelligence.

While our results are able to only demonstrate a single data point, we feel this is significant in that the provisions of the contest allow us to directly compare a group of analysts using a visual analytics toolkit to experts in a particular modeling domain. However, we recognize that this is a far cry from definitively validating the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely statistical solution. Overall, this work points for the need of better methods for evaluating the impact of visual analytics when used for complex problems such as prediction. There are a variety of factors and variables that need to be addressed and controlled, including the level of expertise and the types of visualizations provided. With our current system in place, we have been collecting streaming movie data in a manner similar to the VAST Box Office Challenge and plan to run a variety of controlled experiments. Of primary interest are exploring levels of expertise and the impact that visual analytics has on resultant predictions. We feel that results shown in this paper provide an important starting point for such explorations.

6 Sidebar: Linear Regression Model Construction and Evaluation

Regression analysis is one of the most widely used methods of pattern detection and multifactor analysis [7]. With a proper regression model, data can be better described, interpreted, and predicted.

6.1 Linear Regression Model

The basic form of a k-variable linear regression model is defined as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{4}$$

Variable y is known as the response, variables $x_i, i = 1, \ldots, k$ are the regressors and ε represents the error term. The goal is to define a relationship between the response term and the regressors by solving for the linear coefficients, β_i that best map the regressors to the response. The linear regression model is most often written in matrix form such that:

$$Y = X\boldsymbol{\beta} + \varepsilon$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

For multiple regression models, higher order terms may also be used to model the response (e.g., 2^{nd} order variables are of the form x_i^2 and $x_i x_j$). However, for this work our focus is on the simple linear regression model.

6.2 Parameter Estimation

In order to solve for the parameters β_i the ordinary least square (OLS) solution is most commonly employed. Note that this assumes normality for the data; however, if this assumption is not valid a maximum likelihood estimation would then be employed (which is equivalent to OLS under the assumption of normality).

For OLS, we wish to minimize

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

by satisfying

$$\frac{\partial S}{\partial \boldsymbol{\beta}} \mid_{\hat{\boldsymbol{\beta}}} = -2X^T y + 2X^T X\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

Under assumptions of normality, the solution takes the form of $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$ and the prediction function is $Y = HY$ where $H = X(X^T X)^{-1} X^T$. In one-order multiple linear regression, the predicted response is a linear combination of observations.

6.3 Model Selection

In a multiple variable dataset with a single response variable (such as in our box office gross prediction), analysts will traditionally be faced with a large set of potential linear regression models consisting of various regressors and orders. For example, in box office prediction, the response could be related to the number of Tweets per

day, or the number of theaters the movie is released in, or any combination of variables.

In order to decide which model should be used in prediction, there are several principles an analyst will typically consider.

- Do not violate the scientific principle, if there exists one, behind the dataset.
- Maintain a sense of parsimony to keep the order of the model as low as possible and the number of regressors as small as possible.
- Keep an eye on extrapolation. Regression fits data in a given regressor space but there is no guarantee that the same model also applies to other data outside this space.
- Always check evaluation plots more than the statistics. Residual plots and normal plots help show outliers and lack of fit.

In order to verify the efficacy of a model, analysts will typically rely on a variety of statistical graphics to determine the critical variables in the model, i.e., those that explain the most variation with the simplest form [8]. Several statistics are usually reported to evaluate the effective fit of a given model: p-value, R^2 and R^2-adj. The p-value shows the significance of a regression model, where $p < .05$ indicates the model is significant with a 95% confidence interval. R^2 and R^2-adj generally describe the percentage of variance explained by a given model. R^2-adj specifically takes the degree of freedom into consideration and should be used in multiple regression to compensate for the increased variance when adding regressors. A model is typically selected when it has a small p-value and a high R^2 or R^2-adj value and a relatively simple form with reasonable residual distributions.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. The authors would like to thank the VAST challenge organizers and participants for their help in data collection, evaluation and discussions.

REFERENCES

[1] S. Asur and B. A. Huberman. Predicting the future with social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499, 2010.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[3] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013.

[4] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *IEEE Conference onVisual Analytics Science and Technology (VAST)*, pages 143–152, 2012.

[5] M. C. Hao, C. Rohrdantz, H. Janetzko, D. A. Keim, U. Dayal, L.-E. Haug, M. Hsu, and F. Stoffel. Visual sentiment analysis of customer feedback streams using geo-temporal term associations. *Information Visualization*, 2013.

[6] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296, 2010.

[7] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.

[8] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008.

[10] A. C. Reggie Panaligan. Quantifying movie magic with google search. *Google Whitepaper — Industry Perspectives + User Insights*, 2013.

[11] G. Rowe and G. Wright. The delphi technique as a forecasting tool: Issues and analysis. *International journal of forecasting*, 15(4):353–375, 1999.

[12] T. Schreck and D. Keim. Visual analysis of social media data. *Computer*, 46(5):68–75, 2013.

[13] J. S. Simonoff and I. R. Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000.

[14] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky. I-si: Scalable architecture for analyzing latent topical-level information from social media data. *Computer Graphics Forum*, 31(3pt4):1275–1284, June 2012.

[15] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 301–304, 2009.