Movie Review Comparison on Amazon and Rotten Tomatoes

Elise Thrasher

Illinois Institute of Technology

INTRODUCTION

Although the internet has facilitated the collection and analysis of the general public's ratings and reviews on many topics, data pertaining specifically to movies is currently scattered across several websites which appear to serve the same purpose. This paper attempts to distinguish the uses of two popular rating and review websites: Amazon [1] and Rotten Tomatoes [2].

Due to time constraints, the only features used in this study are the Star Ratings of the movies selected on each site analyzed, and the content of a subset of reviews collected from each site. Additional relevant features that could be incorporated are detailed in the Future Work section of this paper.

Using the data collected directly from the selected websites, I attempt to find a difference in the unigrams, collected from the reviews for each movie, that is able to indicate which site has a higher Star Rating, or if the Star Ratings are similar.

DATA

The data is retrieved from Amazon through the website's category filters. Firstly, the results from Amazon's Movies & TV Department are filtered to Movies, DVD Format, with 1Star & Up. This removes listings for television shows, which are only sparsely listed on Rotten Tomatoes and thus outside the scope of this paper, and cuts down on duplicate listings for the same movie by restricting the formats presented. By filtering by Star Rating, it guarantees that the movies present have some ratings, as the lowest rating a user can give is 1 Star. The movies are further filtered by the program to have a number of reviews over a user-provided threshold, to ensure a significant number of user inputs for analysis.

Movies are further categorized by their release date: Coming Soon (not yet released), Last 90 Days, Last Year (within a year of a user-provided date), and Last Decade (since 2000, but before the date provided for Last Year); none of the movies selected overlap between these categories.

For each category, a user-specified number of movies are selected, allowing or a greater number of reviews collected in the larger release date categories. In this experiment, 50 movies which adhere to the above filters were selected out of the Coming Soon category, 50 movies were selected in Last 90 Days, 100 movies were selected in Last Year, and 300 movies were selected out of Last Decade.

Some movies could not be found on Rotten Tomatoes using the information collected from Amazon. Most frequently, this was due to the specific movie simply not being listed on Rotten Tomatoes for some reason. I can only assume the movies were not popular enough to collect information on. A few movies were listed under different names, abbreviated titles, or contained secondary titles that were not present in the Amazon listing. Lastly, Rotten Tomatoes had several listings which contained a numeric string in the movie's URL that I was unable to decipher satisfactorily to find the movie. All of these factors required that the Rotten Tomatoes listing not be included in the data, as the page containing the reviews could not be found. As such, the Amazon listing was also removed from analysis. As of the data collection on November 26, 2013, this resulted in having 29 listings (out of 50) in the Coming Soon collection, 36 listings (out of 50) in the Last 90 days collection, 78 listings (out of 100) in the Last Year collection, and 176 listings (out of 300) in the Last Decade collection

Reviews and Star Ratings were extracted from both Rotten Tomatoes and Amazon`s review page for each movie selected with the use of the JSoup Java package [3]. The release date of the movie is compared across sites to ensure the correct movie has been found.

METHODS

The most general Star Ratings of each found movie were compared across our sites, adjusting Amazon's ratings to a 10-point system. The value derived by subtracting one star value from the other was condensed to a label: "Amazon" if the Amazon listing was one point or more greater, "RotTom" if the Rotten Tomatoes general rating was greater, or "Balanced" if the rating was less than one point of difference between the two.

The reviews of each movie were simplified into the frequency counts of each unigram. For each site's reviews, the entire text was divided into single words, made lower case and with all non-alphanumeric characters removed. Each word was then counted for each movie on each site. Once the processing was completed, the frequency count of each word was compared across the movie's sites. If the difference in frequency was greater than 5, that is, if the word occurred at least 5 times more in Amazon reviews than in Rotten Tomatoes or vice-versa, the word was included as a feature with the value of the frequency difference.

EXPERIMENTS

The data collected, unigram frequencies and star ratings, were output in an ARFF file format to be analyzed through Weka[4].

Each release date category was output to its own file to be analyzed separately. It is worth noting that, as the Coming Soon and Last 90 Days categories had so few entries, the results from this data is not particularly viable or relevant to future study.

For each ARFF file, the Weka program attempted to classify the movies as having a higher Star Rating on Amazon, a higher rating on Rotten Tomatoes, or having a balanced rating. Through this division, though, it was evident that the original premise of this experiment had to be modified. I had originally assumed that a good number of movies would be assigned the Amazon and Rotten Tomatoes labels, with a few Balanced labels filling out the collection. Instead, it was obvious that Amazon listings had a consistently better rating than Rotten Tomatoes. In fact, only a single entry in the collected data set was labeled as having a significantly higher Rotten Tomatoes rating than Amazon rating (This Is the End (2013)), all others had either an Amazon bias or were balanced, although an Amazon bias was much more common.

As well as the category-specific files, an additional ARFF file was created that combined all the movie listings regardless of release date. In this file, the number of Amazon- and Balanced-labeled movies was equal, removing the frequency bias. This data file contained only 109 entries.

RELATED WORK

There is not a great deal of published work relevant to the topics studied for this paper. Although it is easy to find works which use movie reviews from Amazon or Rotten Tomatoes or similar sources, most of the studies I was able to find were working on review summarization and/or categorization or sentiment analysis. Even if these topics were not directly related to the task at hand, they still proved helpful in formulating my approach to this project.

Zhang et al.'s paper on Movie Review Mining and Summarization [5] collected numerous reviews from Amazon as well as from IMDb (Internet Movie Database) [6] and attempted to summarize the movie's reviews by focusing on film-specific words and opinion words. Once the desired words were identified within the review, the polarity of the opinion word and the topic of the feature word were first extrapolated, then paired. These pairs were then output as the summary.

Although Zhang et al.'s paper provided an interesting take on summarization and identifying opinion, only the sparse data mining information proved practical, with other sections merely providing interesting inspiration for future work.

Wong et al.'s work compares the ratings of movies on IMDb and Rotten Tomatoes with those extrapolated from Twitter using sentiment analysis [7]. Again, the most relevant information is in the Data Mining methods describing how the data was collected and filtered and adjusted for relevancy.

CONCLUSIONS

Not much can be decisively concluded about my original hypothesis, as one of my main assumptions proved inaccurate. That there was a distinctly insignificant number of movies with higher ratings at Rotten Tomatoes over Amazon required me to abandon my original comparison. In its stead, I altered my task to evaluate the unigrams present in Amazon-favored reviews with those that had similar ratings across both sites. I expected this adjustment to provide much less definite results than my original experiment, due to what I would expect to be a smaller difference between the content Amazon and Balanced reviews. The results from Naïve Bayes classification on each category can be viewed in Figures 1, 2, 3, and 4.

	Precision	**Recall**
Amazon	0.931	1
Balanced	0	0
RotTom	0	0

Figure 1: Results from Naïve Bayes classification on 29 Coming Soon movies

	Precision	**Recall**
Amazon	0.806	0.893
Balanced	0.4	0.286
RotTom	0	0

Figure 2: Results from Naïve Bayes classification on 36 Last 90 Days movies

	Precision	**Recall**
Amazon	0.729	0.729
Balanced	0.158	0.167
RotTom	0	0

Figure 3: Results from Naïve Bayes classification on 78 Last Year movies

	Precision	**Recall**
Amazon	0.828	0.776
Balanced	0.238	0.303
RotTom	0	0

Figure 4: Results from Naïve Bayes classification on 176 Last Decade movies

The results are further obfuscated by the sheer number of Amazon reviews in comparison to those in the Balanced category. For example, in the Last Decade dataset used in testing, out of 174 movies, 143 (81.25%) were categorized as Amazon. The other categories had similar biases: Coming Soon had 27 Amazon-labeled entries out of 29 (93.10%), Last 90 Days had 28 out of 36 (77.78%), and Last Year had 59 out of 78 (75.64%). The effects of this bias can be most easily seen in the Coming Soon data analysis as all movies were classified as "Amazon."

Because of the classification bias, where items would be more likely to be labeled as Amazon simply by chance, the release date categories were merged into another file for analysis that contained the same number of Amazon entries as Balanced (See Figure 5 for results). However, the variety in age of these may hinder the results obtained as newer works are more likely to have fewer, less-helpful reviews than those which have stabilized at a high popularity. Nonetheless, this does reduce the classification problem to 50/50 chances and allows the analysis to proceed based solely on the work's features.

	Precision	**Recall**
Amazon	0.66	0.611
Balanced	0.627	0.685
RotTom	0	0

Figure 5: Results from Naïve Bayes classification on unbiased set with data from all categories

As is evident from figure 5, there is a slight improvement from the 50% chance one would expect.

Although I do not know the precise cause of the Amazon bias, it is likely due to the main function of the site dealing with the purchase of movies, rather than simply their rating as Rotten Tomatoes appears to feature. I would expect the majority of the users reviewing movies on Amazon to possess the movie which they are commenting on, and thus rationalize the movie as being of a higher quality than those which they did not own, incorporating a kind of purchaser's bias. Of course, proof of this hypothesis lies firmly in future work and is discussed more in the next section.

As for the features that proved most indicative of the difference between the two analyzed categories, the chart in figure 6 provides the mean and standard deviation for the words with a significant (more than 10 points) difference. The negative mean simply implies that the value was greater in Balanced-labeled movie reviews. To be more specific, for balanced movies, the word "the" occurred more frequently in the Amazon review than in the Rotten Tomatoes review, and the mean difference between these frequency values was greater by 44 points than the comparison of frequency of the same word in Amazon-labeled movies.

Word	**Mean**	**Std Dev**
the	-44.7757	-3.3411
of	-21.1224	-1.6779
and	-17.0677	-3.0357
a	-16.8051	-1.5642
to	-16.3358	0.8659
in	-12.8748	-2.9216

Figure 6: Features from unbiased set with Mean > 10 or Mean < -10

FUTURE WORK

The first step I would recommend future researchers take would be to expand the number of movies collected and/or the number of reviews analyzed. This will ideally provide more data and possibly allow more definite claims to be made without the Amazon bias currently present in this paper's data.

In an effort to increase the data collected and provide more relevant parameters on which to make claims, it may be prudent to introduce more features as well. Data collected on the movie or on the reviewer may hold some clue as to the user's motivations in writing the review. Features like the movie's theater release date, home video release date, box office earnings, total number of reviews per site, and perhaps even the production company and the featured actors or director could be added in to further quantify the relationship between the movie and the user. In addition, reviewer data, like the review's Star Rating, date posted, whether the author is a professional critic (on Rotten Tomatoes) or whether the review is marked as "helpful" by many users (on Amazon), the format reviewed, and the number and quality of the author's previous reviews could also highlight features that quantify the relationship between the reviewer and the movie.

There are several other movie review sites that are not included in this study that may highlight other features: IMDb [6], Fandango [8], Metacritic [9], and MRQE [10] could all be included in future studies to further increase the scope of the analysis.

An option to avoid drastically increasing the number of features analyzed with the increase in movies or review sites would be to group the words into categories. Possibly using WordNet [11] to find the hypernyms and/or the part of speech of the words present in the sites' reviews could greatly decrease computation time, decrease movie-specific topic influence, and produce more accurate results.

SOURCES

1. *Amazon*. Accessed November 26, 2013. http://www.amazon.com/
2. *RottenTomatoes*. Accessed November 26, 2013. http://www.rottentomatoes.com/
3. *JSoup.* Accessed October 14, 2013. http://jsoup.org/
4. *Weka*. Accessed October 25, 2013. http://www.cs.waikato.ac.nz/ml/weka/
5. Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. "Movie review mining and summarization." Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006.
6. *IMDb*. Accessed October 20, 2013. http://www.imdb.com/
7. Wong, Felix Ming Fai, Soumya Sen, and Mung Chiang."Why watching movie tweets won't tell the whole story?." Proceedings of the 2012 ACM Workshop on online social networks. ACM, 2012.
8. *Fandango*. Accessed October 20, 2013. http://www.fandango.com/
9. *Metacritic*. Accessed October 20, 2013. http://www.metacritic.com/movie/
10. *MRQE*. Accessed October 20, 2013. http://www.mrqe.com/
11. *WordNet*. Accessed November 26, 2013. http://wordnet.princeton.edu/