The representation of top critics on movie reviews

s1044478
Radboud University
Nijmegen, Netherlands

ABSTRACT

With the growing numbers of movie reviews online, distinguishing between the perspectives of top critics and average viewers has become crucial for consumers and the film industry. In this study, we analyze a dataset of Rotten Tomatoes movie reviews to explore the differences in language usage, sentiment expression, and review evaluation criteria between top critics and non-top critics. Our analysis includes n-gram analysis to identify language patterns and sentiment analysis to investigate the correlation between sentiment, subjectivity, and review scores. We find that top critics tend to offer more nuanced critiques, while average viewers show more enthusiasm and sentiment in their reviews. This research contributes to understanding the dynamics of movie reviews in the digital age and provides insights for both consumers and industry stakeholders.

KEYWORDS

Movie reviews, top critics, n-gram analysis, sentiment analysis

1 INTRODUCTION

Nowadays, there is an enormous amount of data, and it is growing day by day. This is similarly the case with movie reviews. More and more people leave reviews about a movie they have seen, and new movies are brought out every week. These reviews can help people by choosing whether to watch the movie or let them see whether others have the same opinion about the movie. As there are more and more movie reviews, it can become quite difficult to find movie reviews that are actually useful to read. Therefore, it is easier to read movie reviews from top movie critics. But how well do they represent the opinion of the average movie viewer? In this paper, research to that is done with the help of the Rotten Tomatoes movies and critic reviews dataset[4], which keeps track of whether the review is written by a movie critic or not. After some preprocessing, two different analysis are done on the dataset. Firstly, n-gram analysis, to identify topics and differences in topics between top critics and non-top critics. Secondly, sentiment analysis is done on the reviews to see the correlation between sentiment, subjectivity and the review scores. This is done for top critics and non-top critics separately, so that the differences can be seen. This paper consists of a review of the related work on this subject, followed by a description of the methodology. It will then show the experimental results and discuss them.

2 RELATED WORK

N-grams are used to count occurrences of word groups[5]. For instance, if you let n be 2, then a much occurring 2-gram might be "good film". Just like the word count vector, all n-grams in all the reviews will be counted to get features. N-grams can be a good measure for catching the essence of a piece of text. So, I visualized that in a word cloud, to see the essence of a lot of reviews in one image. Hazarika et al.[3] conducted a study on sentiment analysis on Twitter using TextBlob for natural language processing. Their research focused on analyzing sentiment polarity and subjectivity in tweets to understand public opinion and sentiment trends on various topics. By leveraging TextBlob, a Python library for processing textual data, they investigated the effectiveness of sentiment analysis techniques in the context of social media data. Their findings contribute to the broader understanding of sentiment analysis methodologies and their applications in capturing public sentiment in real-time on social media platforms like Twitter.

3 METHOD

3.1 Dataset

For the data, a dataset from Kaggle about Rotten Tomatoes movies and critic reviews is used[4]. This dataset contains 1130017 movie reviews. The reason that this dataset is used, is because it contains a column with information of whether the review is written by a top movie critic or not. This enables a comparison between top critics and non-top critics. The dataset consists of two files. One with reviews and one with information about the movie. For this paper, only the file with the reviews is used and only 4 columns of this file. The columns that are used are: a Boolean value about whether a top critic has written the review, the score of the review and of course the content of the review itself.

3.2 Processing pipeline

The data is loaded and then pre-processed. The pre-processing consist of a few steps. Firstly, the rows with empty values will be removed. Secondly, the scores of the reviews are made consistent. The scores in the dataset vary from numbers, to fractions and even letters. So all of those are converted to a number between 1 and 5. The reviews of the scores that can not be converted are removed. Some of the reviews are written in a different language than English, so these are filtered out as well. Just like the reviews that occur more than once in the dataset. For the n-gram analysis, an extra column is added with the review tokenized. This is done with the Tweet Tokenizer from the NLTK library[1], because some of the movie reviews contain tokens that you can also come across in tweets.

For the n-gram analysis, the n-grams are extracted from the tokenized text data. It utilizes the n-grams function from the NLTK library to generate n-grams of a specified length. Additionally, it filters out n-grams starting with stopwords. Stopwords are common words that may not carry significant meaning. All the n-grams are counted, and visualized in a word cloud. In this word cloud the size of each n-gram is represented by the frequency in the distribution. By visualizing it this way, the most common n-grams can be easily spotted. This is done on two subsets of the data: reviews from top critics and reviews from non-top critics. Each subset is randomly sampled to ensure a consistent sample size of 75,000 reviews for a good comparison between the sets. The word clouds are plotted for multiple n-gram lengths (4, 5, 6, and 7) to analyze the distribution of n-grams across different lengths.

For the sentiment analysis, first the distribution of review scores is visualized to understand the overall rating of the reviews. Then the sentiment and subjectivity of the reviews are calculated by using the TextBlob library[2]. The correlation between these scores and the review score is calculated with the Pearson correlation coefficient for both the top critics and the non-top critics.

The full code implementation of the pipeline can be found on GitHub.[1]

4 RESULTS

4.1 N-gram analysis

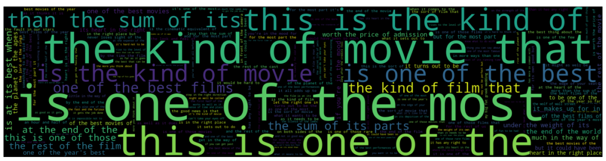

Figure 1: Word cloud of 5-grams from top critic reviews

From figure 1 you can see that the most common 5-grams for the top critic reviews are "is one of the most", "the kind of movie that" and "this is one of the". They respectively have 50, 34 and 30 occurrences.

Figure 2: Word cloud of 5-grams from non-top critic reviews

From figure 2 you can see that the most common 5-grams for the non-top critic reviews are "one of the best films", "of the best films of" and "is one of the best". They respectively have 111, 83 and 72 occurrences.

Figure 3: Word cloud of 6-grams from top critic reviews

From figure 3 you can see that the most common 6-grams for the top critic reviews are "than the sum of its parts", "its heart in the right place" and "one of the best movies of". They respectively have 19, 16 and 16 occurrences.

[1]https://github.com/jveldik/TxMM-project

Figure 4: Word cloud of 6-grams from non-top critic reviews

From figure 4 you can see that the most common 6-grams for the non-top critic reviews are "one of the best films of", "of the best films of the" and "the best films of the year". They respectively have 83, 57 and 55 occurrences.

4.2 Sentiment analysis

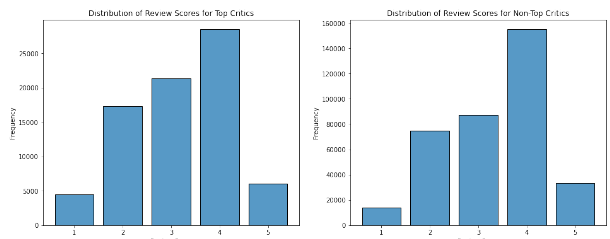

Figure 5: Distribution of review scores

From the score distribution in figure 5 you can see that the top critics are slightly more critical. They give respectively more reviews with lower scores than the non-top critics, who seem to have a higher peak at the review score 4.

Top critic	#Reviews	Review score	Sentiment	Subjectivity
True	77693	3.1838	0.1166	0.4997
False	364312	3.3268	0.1380	0.5109

Table 1: Scores of top critics and non-top critics

Table 1 shows the results of the average scores for both the subsets. The top critics have slightly lower values for all the scores.

Correlation	Top critics	Non-top critics
Sentiment vs. review scores	0.2233	0.2689
Subjectivity vs. review scores	0.0695	0.0683
Sentiment vs. subjectivity scores	0.1592	0.1683

Table 2: Correlations between scores

Table 2 shows the results of the Pearson correlation coefficient between the different scores. The top critics have a slightly lower correlation between the sentiment in their reviews and the review scores. The correlation between the subjectivity and the review scores is almost the same for all reviews. That is also the case for the correlation between the sentiment and the subjectivity scores.

5 DISCUSSION

5.1 N-gram analysis

The n-gram analysis provides an interesting insight into the language patterns characteristic of each group. For instance, the prevalence of certain phrases like "one of the best films" in non-top critic reviews suggests a tendency towards enthusiasm and positivity. On the other hand, top critics might lean towards nuanced critiques and going deeper into the content of the movie, as indicated by phrases like "than the sum of its parts" in their reviews. These differences in language usage hint at varying priorities and perspectives between the two groups.

5.2 Sentiment analysis

By examining the distribution of review scores and calculating sentiment and subjectivity scores, we hoped to shed some light on potential differences in how the two groups evaluate movies. The observation that top critics tend to be slightly more critical, is also visible in their distribution of review scores. The correlation analysis reveals some more nuanced relationships between sentiment, subjectivity, and the review scores. While both groups have similar correlations between subjectivity and review scores, there are subtle differences in the correlation between sentiment and review scores. Notably, top critics have a slightly weaker correlation between sentiment and review scores compared to non-top critics. This suggests that top critics might prioritize aspects beyond emotional feelings when assigning scores, like how well-made it is or if the story makes sense.

6 CONCLUSION

One of the central questions addressed in this paper is how well top critics represent the opinion of the average movie viewer. While top critics have influence and authority in the world of film criticism, their perspectives might not be the same as those of the general audience. This study shows that the average reviewers have more sentiment in their reviews and give on average higher scores, while the top critics have slightly less sentiment in their reviews and also talk more about other aspects of the film. Despite its insightful findings, this study is not without limitations. The dataset's reliance on Rotten Tomatoes reviews might introduce biases to the platform and its reviewers. So, in the future this study can be expanded by making analyzing a dataset from another platform and comparing the results. Another direction for further research could look deeper into the coherence, and complexity of the reviews.

REFERENCES

[1] [n. d.]. NLTK :: nltk.tokenize.casual module. https://www.nltk.org/api/nltk.tokenize.casual.html

[2] [n. d.]. NLTK :: nltk.tokenize.casual module. https://www.nltk.org/api/nltk.tokenize.casual.html

[3] Ditiman Hazarika, Gopal Konwar, Shuvam Deb, and Dibya Jyoti Bora. 2020. Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing. In *International Conference on Research in Management & Technovation*. https://api.semanticscholar.org/CorpusID:234365972

[4] Stefano Leone. 2020. Rotten tomatoes movies and critic Reviews Dataset. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_critic_reviews.csv

[5] Prayag Tiwari, Brojo Kishore Mishra, Sachin Kumar, and Vivek Kumar. 2017. Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis. *International Journal of Knowledge Discovery in Bioinformatics* 7, 1 (Jan. 2017), 30–41. https://doi.org/10.4018/ijkdb.2017010103

A WORK REPORT

I started working in another direction with this report. In my first attempt, I tried to make two models to predict the review scores for both the top critics and the non-top critics. This was not a success, because the models had low accuracies and I realized that it was more of a machine learning approach. So I decided to focus more on the differences between the two groups in the reviews. After some research and looking at assignment 1 to get some inspiration. I rewrote my code to do n-gram analysis. However, I felt that that was not enough for this paper, so I added code for sentiment analysis. This time with an existing library instead of making a model myself. When the code was all done, I started rewriting the report to match the new code I had written. Unfortunately, I started too late to finish the related work section.