NLP Comparison between Movie Reviews from Critics and Consumers

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Yuelan Sheng

Spring, 2020.

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Yangfeng Ji, Department of Computer Science

Technical Topic - NLP Comparison between Movie Reviews from Critics and Consumers

What insights and marketing strategies can be found by comparing the actual texts between professional critics and normal consumers using natural language processing?

Introduction and Objectives

Upon entering a new era in which big data and algorithms will revolutionize people's lives, people start to pay a significant amount of attention to how data being collected over the last few decades can be used. Thus, there are more and more studies applying machine learning techniques in developing better marketing strategies for different industries. In the movie industry, in order to improve the box office sales, researchers and companies have been studying the potential factors that affect the sales of the movies. The past studies have been mainly focused on the characteristics of the movies, such as genres, directors, leading actors/actresses, and main themes. Some of the studies also apply natural language processing (NLP) techniques in figuring out how the written reviews of the movies can affect the box office performance, which mostly studied the sentiments and named-entities in the reviews. These studies have not separated reviews into different discussions. In reality, there are two major groups of movie reviewers: professional critics and regular consumers. These two groups are highly possible to have different perspectives and emphasis on evaluating the movies. Since these two groups of reviews focus on their own points, ultimately creating reviews of different effects. There is huge marketing potential in applying the differences of reviews since the reviews can be promoted in a certain way that influences people's decisions of going to a movie or not, and the information of the reviews can be used as a model for predictions of the box office of a certain movie.

In order to better understand the differences in movie reviews from the two distinct groups and relate the differences to effects on box office sales, sentiment analysis, topic modeling, and other statistical analysis need to be conducted on the dataset. The results found in the statistical and NLP analysis will be helpful in introducing new theories to explain consumer activities and marketing strategies in the movie industry.

Data Collection

The consumer review data is obtained by scanning through more than 10,000 movies on the IMDb website and scrapping the consumers' reviews from IMDb. Then the reviews from professional critics are scrapped using top-10 film review websites: the New York Times, San Francisco Chronicle, Roger Ebert, The Onion AV Club, Slant Magazine, eFilmCritic, Blu-Ray, Austin Chronicle, Pop Matters, and Common Sense Media. In order to eliminate potential biases, movies that do not have at least three reviews from both the critics' reviews pool and the users' pool have been removed from the master dataset. The metadata of each movie is collected through the description of IMDb, including genres, directors, leading characters, production companies, oscar nomination, movie runtime, number of raters, overall ratings, and final sales. The metadata serves as an aid to better explain how differences in critics' and users' reviews are related to movies.

Methodology and Procedures

Before applying any NLP techniques, it is important to perform some data preprocessing steps to better facilitate the following discussions. Primarily, evidence-free analysis using statistical mechanics is performed in order to learn about whether the underlying assumptions hold. The two assumptions are the two scores from critics' ratings and users' ratings are not highly correlated, and the scores from critics and users are highly correlated to the final box office sales. Different regression models and correlation matrices are trained on the statistics of the reviews with box office sales, in order to see the correlations among characteristics of movie reviews and box office sales. Secondly, the total number of reviews is over 1.4 million, which means there can be hidden insights if the reviews are divided into smaller groups. Since movies are clearly defined with genres and each genre of movies has specific highlights in production, the original dataset is separated into seven main genres: action, animation, comedy, drama, horror, sci-fi, and others.

For the NLP analysis, the two main focuses on this project are to learn about sentiments and topics of each review. Sentiment analysis refers to the idea of systematically identify the emotions behind a review, generally giving a score that can represent positive, negative, and neutral sentiments. To obtain the topics of reviews, it is essential to put the texts of all reviews into one corpus and perform LDA training. The process of LDA modeling is unsupervised and from the LDA model, keywords of each topic can be extracted and used for further analysis.

After performing sentiment analysis on the dataset, the result will give a general pattern of how a certain group of reviewers, either professional critics or normal consumers, prefers to talk about the movie. Two different packages are used in the analysis, NLTK and StanfordCoreNLP. NLTK is used to generate one sentiment score for each individual review, so the overall emotion can be detected. StanfordCoreNLP performs sentence to sentence sentiment analysis; therefore, the transitions and variations of emotions in the reviews can be found. Additionally, the general features of reviews are also calculated, such as lengths, named entities, and word frequencies.

The next step in NLP analysis is to figure out what specifically the movie reviews talk about; therefore, topic modeling is performed on all the effective reviews so it is possible to conclude about what are the topics the people generally include when evaluating a certain movie. Furthermore, since the movie reviews are divided into different genres, it is possible to conclude information about specific areas. The LDA model is set to have an alpha value of 0.1 and a number of topics of 5. For each genre, the reviews are trained to five topics and the top thirty words of each topic will be extracted. Then, each topic is assigned with a self-explanation that summarizes what this subject of reviews mostly talks about. Also, since there is meta-information about whether the review is from a critic or not, it is possible to generate distributions of reviews from different groups of reviewers. Additionally, since the dataset is too large to be handled with an existing package, the LDA model was chosen for this project is the online LDA model developed by Blei Lab ((Blei-Lab, 2016)).

The above two steps are preliminary preparation for the more sophisticated model of topic-specific sentiment analysis. The objective of topic-specific sentiment analysis is to help understand whether there is a heavier weight on a certain topic that contributes to the overall rating of a movie. The calculated results of sentiment scores for different topics within one review are applied to machine learning models to calculate how some topics may be more important in concluding an overall sentiment score for a movie.

Results and Discussion

For the data-preprocessing and evidence-free analysis, it is found that the dataset has met all the assumptions. First of all, the critics' ratings and users' ratings are not highly correlated, meaning that it is reasonable to study them separately. Secondly, the ratings from both groups are highly correlated to the overall rating and box office of the movies, which indicates the logic of studying how reviews affect the performance of a movie. Moreover, the other data exploration steps of lengths and named entities show that critics generally have longer reviews than the normal consumers, and critics tend to use a lot of named entities in the reviews, which are some professional vocabulary. Therefore, one hypothesis is made that critics generally write longer reviews because they tend to talk about more aspects of the movies, requiring more professional terms used in the reviews.

The result of preliminary sentiment analysis has shown that for most movies, critics tend to have similar sentiments in commenting, meaning that most of them agree with one and another. However, the variations of sentiments in regular users' reviews are a lot greater than in critics'

reviews. This leads to the conclusion that when users are making judgement of a movie, they seem to be very personal and thus cannot discuss every aspect of the movie fairly, while the critics are more professional and subjective since they consider more than just one aspect and will not be biased for most parts.

In LDA topic modeling, it is shown that each genre has its own focuses on topics but at the same time, they have similar topics. For example, across all different genres, topics containing words that describe the actors/actresses and directors are present, showing that the production team is still an essential topic in discussing the quality of the movies. Also, for each genre, the topics modeling distributions have shown that each genre will have a few sub-genre within it that people like to discuss the most, such as the romantic comedy under comedy or alien story under sci-fi. With the model trained by LDA, it is possible to learn about the distributions of topics each review has talked about depending on whether it is from a professional critic or a regular consumer. Using the resulting distributions, the calculated entropies can indicate a general pattern of how reviewers write comments. Entropies for distributions of critics are mostly greater than or equal to 2, meaning that the distributions are more uniform while the distributions of users are smaller, meaning that the distributions are only in some topics. The result has shown that the hypothesis brought up in the previous evidence-free analysis is correct that most reviews from critics discuss evenly among the topics generated by the online LDA model while on the other hand, regular users tend to only discuss one or two topics.

Conclusion and Further Research

The massive amount of online reviews has proven that critics and users do have their own perspectives in writing reviews. Users are more likely to be focused on one specific area in the movie. For example, when watching an action movie, the audience cares most about the actual action scenes in the movies and if they enjoy the action part, they may just rate the movie as a great movie; however, the critics not only care about the action scenes, they also care about the production teams and other logics in the stories. Therefore, the sentiments specified with each topic are also important factors in predicting the effects of movie reviews on the overall ratings and total box office.

In a future study, it is worthwhile studying the effects of reviews over time with the time to time box office sales, which is to get a dynamic model in analyzing the reviews and ticket sales. Moreover, it is possible to apply advanced models such as the BERT model to calculate the aspect-specific sentiments for each sentence in the reviews and get some additional insights ((Jiang, Chen, Xu, Ao, & Yang, 2019)).

Reference

Blei-Lab. (2016, October 11). blei-lab/onlineldavb. Retrieved from https://github.com/blei-lab/onlineldavb

Jiang, Q., Chen, L., Xu, R., Ao, X., & Yang, M. (2019). A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

doi: 10.18653/v1/d19-1654