# How do Reviews Impact Movies?


***BSAN-6200-Section-01****: Text Mining & Social Media Analytics*

Author(s):
Jonathan Ting, Krissy Wong

05/01/2022

# Table of Contents

# Background

### Objective

The objective of this report is to showcase what types of aspects in a film should production companies/film makers consider when trying to create a film that generates high reviews on IMDB.

### Business and problem overview/project summary

We hope to gather useful insights/information that will be pertinent to filmmakers or production companies because according to a report over 80% of moviegoers look at a review before attending a film in person. The importance of this is that if a film has a negatively associated review then they will be less inclined to watch it and also vice versa with positive reviews equating to more live viewership.

Do you read reviews before deciding whether to watch a film?

**Open thread: A new report has indicated that eight out of 10 moviegoers in the US refer to film critics before deciding whether to part with their cash. Does that sound like you?**

How do you decide whether to stump up for a cinema ticket? Photograph: PhotosIndia.com

# Data processing

### Data Collection

Our group received the IMDB dataset from Dr. Zaman for our project. The dataset contained the columns for movie title, dates, useful votes, stars, total stars, and reviews. We also wanted to analyze texts based on reviews

from movies from different genres. We collected data on all types of film (eg. movie, short film, TV series, TV episodes, video, etc. ) from IMDB website. The dataset included the following columns: "unique identifier", "type", "primary title", "original title", "Adult" to identify if the film is an adult film, "start year", "end year", "runtime minutes", and "genres". The film dataset from IMDB contained data from 1930 to present day.

## Data Cleaning

The IMDB film genre dataset was large and contained over five million rows of data. It would be inefficient to directly upload a dataset of this size onto Google Colab since it would most likely crash while uploading the dataset. We uploaded the film dataset into Tableau Prep to rename column headings and delete columns not needed for our project. We were only interested in keeping the "primary title", "type", and "genre" columns. Additionally, we filtered the dataset to only contain information on movies. After filtering, we had roughly one million rows for the movie genre dataset. We did a left join on the IMDB dataset with the movie genre dataset on the movie title. Below is the code used to merge the two datasets:

```
# Merge imdb_reviews with movie_genre

# rows in amazon_reviews and team2 with the same text
clean_imdb_reviews = pd.merge(imdb_reviews, movie_genre, how='left', on = 'Movie')

clean_imdb_reviews
```

| | Movie | Date | Stars | Useful Votes | Total Votes | Review | Type | Genre |
|---|---|---|---|---|---|---|---|---|
| 0 | Inception | 8/22/2010 | 9 | 1572.0 | 1813.0 | I'd like to keep my review rather to the point... | movie | Action,Adventure,Sci-Fi |
| 1 | Inception | 7/10/2010 | 10 | 1770.0 | 2568.0 | What is the most resilient parasite? An Idea! ... | movie | Action,Adventure,Sci-Fi |
| 2 | Inception | 7/12/2010 | 10 | 1401.0 | 2303.0 | Usually I try to be careful with over hyping a... | movie | Action,Adventure,Sci-Fi |
| 3 | Inception | 7/13/2010 | 10 | 1150.0 | 1825.0 | Films about dreams and the subconscious are us... | movie | Action,Adventure,Sci-Fi |
| 4 | Inception | 7/9/2010 | 10 | 1102.0 | 1780.0 | I saw Memento very recently, something that tu... | movie | Action,Adventure,Sci-Fi |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 372795 | Killer Elite | 2/19/2015 | 6 | 3.0 | 3.0 | Stirring as well as non-stop action movie , be... | movie | Action,Crime,Thriller |
| 372796 | Killer Elite | 1/15/2012 | 3 | 9.0 | 15.0 | Should have been so much better than this. The... | movie | Action,Crime,Thriller |
| 372797 | Killer Elite | 9/30/2012 | 2 | 14.0 | 25.0 | OK movie in a cut-out type of entertainment. Y... | movie | Action,Crime,Thriller |
| 372798 | Killer Elite | 9/27/2011 | 8 | 14.0 | 25.0 | This movie was far more textured and interesti... | movie | Action,Crime,Thriller |
| 372799 | Killer Elite | 5/18/2014 | 6 | 3.0 | 4.0 | Based on a true story?? It's 1980. Danny (Jaso... | movie | Action,Crime,Thriller |

After merging the IMDB dataset with the movie genre dataset, we noticed there are movies with multiple genres. For example, "Inception" is classified as an Action, Adventure, and Sci-Fi movie. To simplify, we wanted to limit each movie to one genre. We exported the clean_imdb_reviews into an Excel spreadsheet and used Text to Columns to separate each genre under the column "Genre" so that each column contained one genre. The first genre listed for each movie is used as its primary genre. As shown in the screenshots below, instead of belonging to three genre categories, the movie "Inception" is listed as an Action film.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Movie | Date | Stars | Useful Vot | Total Vote | Review | Type | Genre | |
| 2 | 237475 | Inception | ######## | 10 | 1 | 9 | The best | movie | Action, Drama, Sci Fi | |
| 3 | 26303 | Noah | ######## | 2 | 10 | 20 | As widely | movie | Drama, Historical | |
| 4 | 18076 | The Intervi | ######## | 1 | 10 | 22 | If you like | movie | Documentary, Comedy | |
| 5 | 215387 | Drive | 4/4/2012 | 10 | 1 | 3 | I | movie | Documentary, Action | |

| | Movie | Date | Stars | Useful Vot | Total Vote | Review | Type | Genre |
|---|---|---|---|---|---|---|---|---|
| 237475 | Inception | 8/11/2010 | 10 | 1 | 9 | The best | movie | Action |
| 26303 | Noah | 4/21/2014 | 2 | 10 | 20 | As widely | movie | Drama |
| 18076 | The Interview | 12/26/2014 | 1 | 10 | 22 | If you like | movie | Comedy |
| 215387 | Drive | 4/4/2012 | 10 | 1 | 3 | I | movie | Documentary |

Next, we uploaded the updated clean_imdb_reviews dataset to Google Colab and conducted exploratory data analysis (EDA) to identify errors, null values, and unique values in our dataset. Listed is our EDA process:

- Describe the dataset

```
clean_imdb_reviews.describe()
```

| | Stars | Useful Votes | Total Votes |
|---|---|---|---|
| count | 372800.000000 | 372752.000000 | 372752.000000 |
| mean | 6.438476 | 9.956089 | 19.377189 |
| std | 2.861723 | 43.803893 | 71.780640 |
| min | 1.000000 | 0.000000 | 0.000000 |
| 25% | 4.000000 | 1.000000 | 2.000000 |
| 50% | 7.000000 | 2.000000 | 6.000000 |
| 75% | 9.000000 | 6.000000 | 13.000000 |
| max | 10.000000 | 3716.000000 | 5610.000000 |

- Count the number of rows in each column

```
[ ] clean_imdb_reviews.count()

Movie           372800
Date            372800
Stars           372800
Useful Votes    372752
Total Votes     372752
Review          372800
Type            334384
Genre           334384
dtype: int64
```

- Look at the shape of the dataset

```
[ ]  clean_imdb_reviews.shape

     (372800, 8)
```

- Count the total number of null values in each column

```
[ ]  clean_imdb_reviews.isnull().sum()

     Movie                0
     Date                 0
     Stars                0
     Useful Votes        48
     Total Votes         48
     Review               0
     Type             38416
     Genre            38416
     dtype: int64
```

- Count the number of unique values in each column

```
     print(clean_imdb_reviews.nunique())

     Movie              320
     Date              2467
     Stars               10
     Useful Votes       670
     Total Votes        947
     Review          170935
     Type                 1
     Genre              169
     dtype: int64
```

- Get a concise summary of the dataset

```
clean_imdb_reviews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 372800 entries, 0 to 372799
Data columns (total 8 columns):
 #   Column        Non-Null Count   Dtype  
---  ------        --------------   -----  
 0   Movie         372800 non-null  object 
 1   Date          372800 non-null  object 
 2   Stars         372800 non-null  int64  
 3   Useful Votes  372752 non-null  float64
 4   Total Votes   372752 non-null  float64
 5   Review        372800 non-null  object 
 6   Type          334384 non-null  object 
 7   Genre         334384 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 25.6+ MB
```

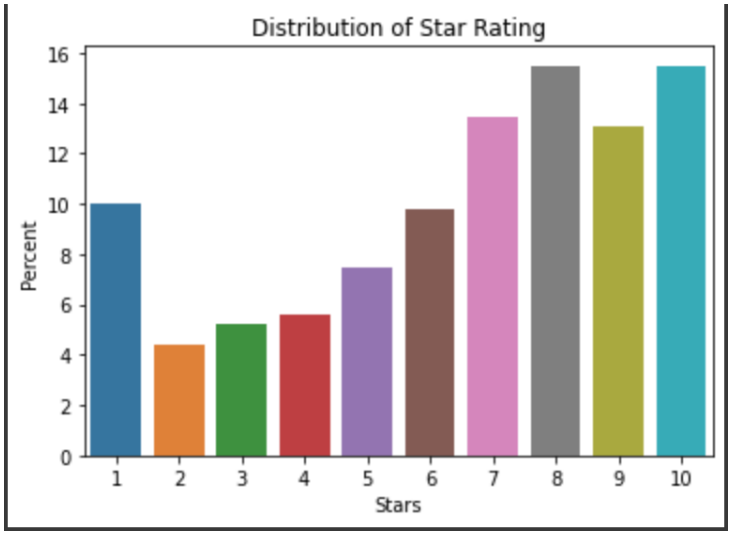
We learned that the clean_imdb_reviews dataset was too large to run analysis and would often crash, so we decided to use a sample of 40,000 rows, which represents about 12% of the dataset. We used the RAND function to conduct a simple random sample technique to extract 40,000 rows for our sample dataset.

# Data analysis

### Specifics

Our group used natural language processing techniques to analyze how imdb reviews impact star ratings. We used WordClouds, Sentiment Analysis, and Deep Learning analysis using BERT classifier, FastText, and Glove2Vec. First, we looked at the distribution reviews among star ratings and saw that there was a polarizing number of reviews between high ratings and low ratings. Movies with star ratings of 1, 7, 8, 9, and 10 were more likely to have reviews. This observation shows that people are more likely to review movies when they perceive the movie to be terrible or amazing.

```
# Use bar plot to visualize distributions of star ratings
ax = sns.barplot(data=clean_imdb_reviews, x='Stars', y='Stars', estimator=lambda x: len(x) 
ax.set(ylabel="Percent")
plt.title('Distribution of Star Rating')
plt.show()
```



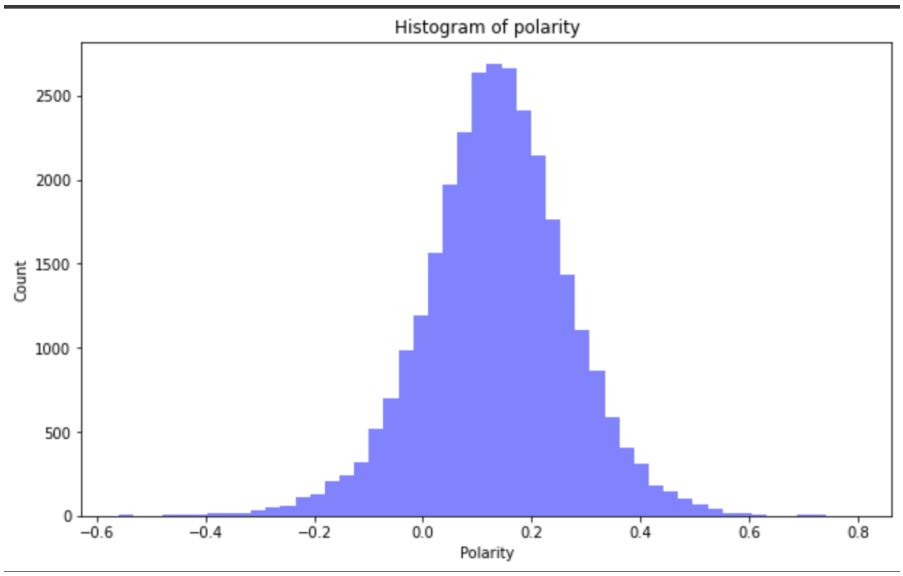

We created a new dataframe to filter for 10 star and 1 star reviews to create WordClouds of most common words associated with reviews with 1 star and 10 star ratings.

Common words from reviews with low ratings (1 star)

Common words from reviews with high ratings (10 star)


As shown in the WordClouds, there are overlapping words in both reviews with low and high star ratings. The most common words associated with low and high-ratings reviews are “even”, “best”, and “think”. Since these words appeared most frequently in both types of reviews, it hints that people were more objective when they wrote reviews.

Next, we conducted a sentiment analysis to determine whether the overall movie reviews have a positive, negative, or neutral sentiment. Since the imdb movies reviews dataset contained over 300,000 rows, we decided to create a sample of 30,000 rows, which represent 10% of the dataset. We created new columns for sentiment and subjectivity, then we created a histogram to visualize the distribution. The range of polarity from our histogram ranges from -0.4 to 0.6. Polarity ranges from -1 to 1, with -1 meaning negative sentiment and 1 meaning positive sentiment. Our histogram is slightly skewed towards positive one, meaning that generally, there is more positive sentiment among movie reviews.



For our final analysis, we used deep learning models, specifically FastText, word2vec, GLOVE, and BERT. Our classification report for our BERT model shows that the weighted average for precision, recall, and f1 score is around 20%. The results are strong since our model predicts using ten columns: Movie, Date, Stars, Useful Votes, Total Votes, Review, Type, Genre, sentiment, and subjectivity. In the next section, we conducted further research on movie reviews in different genres using JMP analysis.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.29 | 0.64 | 0.40 | 754 |
| 2 | 0.00 | 0.00 | 0.00 | 321 |
| 3 | 0.22 | 0.00 | 0.01 | 412 |
| 4 | 0.20 | 0.00 | 0.01 | 435 |
| 5 | 0.19 | 0.02 | 0.04 | 538 |
| 6 | 0.13 | 0.01 | 0.02 | 727 |
| 7 | 0.21 | 0.32 | 0.25 | 991 |
| 8 | 0.22 | 0.41 | 0.28 | 1163 |
| 9 | 0.21 | 0.10 | 0.13 | 1012 |
| 10 | 0.35 | 0.45 | 0.39 | 1147 |
| accuracy | | | 0.26 | 7500 |
| macro avg | 0.20 | 0.20 | 0.15 | 7500 |
| weighted avg | 0.22 | 0.26 | 0.20 | 7500 |

## General Patterns

For our WordClouds we generated some on JMP Pro and analyzed the top 3 genres by review count which were Adventure, Comedy, and Drama. We then categorized the reviews as "high rated" which were reviews that gave the film 8-10 stars and "low rated" which gave the film 1-3 stars. With these word clouds we were able to see a very general pattern of what keywords the reviews were suggesting in regard to what makes a film high rated or low rated.



In the above word clouds we can see the word clouds for Action Films with black words being high rated reviews and red words being low rated reviews. We realized there was quite a bit of overlap in these word clouds so we decided to look at the phrase analysis that JMP provided.

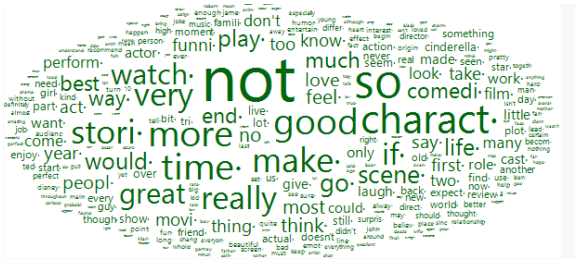
| Phrase | Count | N |
|---|---|---|
| so much | 108 | 2 |
| special effects | 96 | 2 |
| so many | 90 | 2 |
| fantastic four | 85 | 2 |
| ever seen | 81 | 2 |
| don't know | 78 | 2 |
| robin hood | 75 | 2 |
| too much | 75 | 2 |
| waste of time | 69 | 3 |
| if you want | 64 | 3 |
| don't waste | 64 | 2 |
| sci fi | 62 | 2 |
| character development | 60 | 2 |
| so bad | 59 | 2 |
| story line | 56 | 2 |
| not only | 55 | 2 |
| year old | 52 | 2 |
| comic book | 48 | 2 |
| 2 hours | 46 | 2 |
| action scenes | 45 | 2 |
| waste your time | 42 | 3 |
| iron man | 42 | 2 |
| bad guys | 41 | 2 |
| much more | 40 | 2 |
| worst movies | 40 | 2 |

| Phrase | Count |
|---|---|
| dark knight | 284 |
| iron man | 228 |
| special effects | 169 |
| so much | 161 |
| not only | 145 |
| 10 10 | 142 |
| christopher nolan | 142 |
| ever seen | 135 |
| comic book | 130 |
| so many | 130 |
| tom hardy | 130 |
| sci fi | 114 |
| captain america | 110 |
| very good | 110 |
| christian bale | 105 |
| great job | 104 |
| man of steel | 99 |
| much more | 97 |
| dark knight rises | 94 |
| knight rises | 94 |
| action scenes | 92 |
| don't know | 89 |
| big screen | 81 |
| long time | 81 |
| first time | 79 |

In these phrase lists (low on the left, and high on the right) we were able to determine that for Action films, the viewers tend to gravitate towards Superhero films such as "The Dark Knight", "Iron Man", and "Captain America". In regard to the negative aspects, viewers tended to dislike Action films that had poor special effects and a lack of character development/story line.

| Phrase | Count | N |
|---|---|---|
| so much | 443 | 2 |
| very good | 361 | 2 |
| man of steel | 354 | 3 |
| carey mulligan | 348 | 2 |
| sci fi | 344 | 2 |
| not only | 324 | 2 |
| too much | 308 | 2 |
| new york | 305 | 2 |
| so many | 300 | 2 |
| michael fassbender | 293 | 2 |
| special effects | 275 | 2 |
| much more | 262 | 2 |
| don't know | 230 | 2 |
| ever seen | 212 | 2 |
| great job | 199 | 2 |
| iron man | 191 | 2 |
| steve mcqueen | 189 | 2 |
| long time | 185 | 2 |
| love story | 180 | 2 |
| first time | 178 | 2 |
| really good | 174 | 2 |
| over the top | 173 | 3 |
| great gatsby | 173 | 2 |
| ryan gosling | 173 | 2 |
| amy adams | 160 | 2 |

| Phrase | Count | N |
|---|---|---|
| so much | 211 | 2 |
| too much | 183 | 2 |
| special effects | 177 | 2 |
| don't know | 163 | 2 |
| so many | 155 | 2 |
| ever seen | 132 | 2 |
| new york | 131 | 2 |
| man of steel | 121 | 3 |
| sci fi | 121 | 2 |
| year old | 113 | 2 |
| much more | 106 | 2 |
| character development | 105 | 2 |
| much better | 100 | 2 |
| would not | 97 | 2 |
| pretty much | 93 | 2 |
| not only | 92 | 2 |
| 2 hours | 89 | 2 |
| if you want | 88 | 3 |
| waste of time | 87 | 3 |
| very good | 87 | 2 |
| could not | 86 | 2 |
| main character | 83 | 2 |
| over the top | 82 | 3 |
| too many | 82 | 2 |
| too long | 79 | 2 |

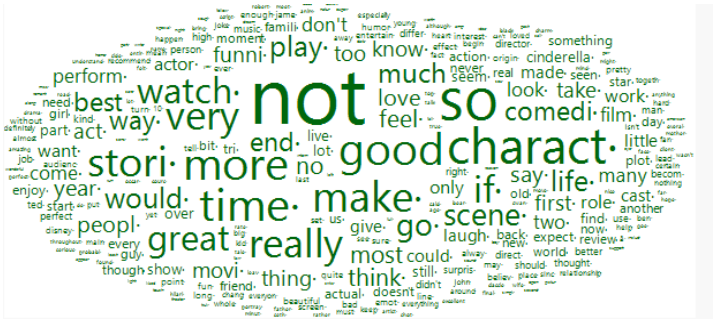




For the Drama category, we could see that there wasn't really a general trend and that very common sentiments across all the other categories were quite similar. This may be due to the fact that the Drama category has a mix of all different types of genre such as sci-fi, comedy, action, documentary, etc. In a sense we treated Drama as a more general category but for future research we would redefine Drama and pick a unique archetype that fits Drama via feature engineering.

| Phrase | Count | N |
|---|---|---|
| so much | 78 | 2 |
| too much | 75 | 2 |
| year old | 67 | 2 |
| don't know | 63 | 2 |
| so many | 61 | 2 |
| family guy | 53 | 2 |
| don't waste | 51 | 2 |
| pretty much | 51 | 2 |
| over the top | 49 | 3 |
| ever seen | 48 | 2 |
| adam sandler | 46 | 2 |
| main character | 46 | 2 |
| seth rogen | 46 | 2 |
| much more | 44 | 2 |
| black swan | 43 | 2 |
| not funny | 43 | 2 |
| would not | 41 | 2 |
| waste of time | 40 | 3 |
| grown ups | 40 | 2 |
| main characters | 39 | 2 |
| much better | 39 | 2 |
| natalie portman | 39 | 2 |
| 2 hours | 37 | 2 |
| could not | 36 | 2 |
| don't think | 36 | 2 |

| Phrase | Count | N |
|---|---|---|
| so much | 159 | 2 |
| de niro | 145 | 2 |
| black swan | 139 | 2 |
| year old | 123 | 2 |
| very good | 114 | 2 |
| so many | 112 | 2 |
| not only | 106 | 2 |
| too much | 106 | 2 |
| lily james | 102 | 2 |
| much more | 97 | 2 |
| live action | 96 | 2 |
| mila kunis | 92 | 2 |
| real life | 87 | 2 |
| teddy bear | 87 | 2 |
| don't know | 86 | 2 |
| family guy | 86 | 2 |
| anne hathaway | 82 | 2 |
| cate blanchett | 81 | 2 |
| fairy tale | 81 | 2 |
| great job | 79 | 2 |
| special effects | 79 | 2 |
| robert de | 78 | 2 |
| side effects | 78 | 2 |
| natalie portman | 77 | 2 |
| very funny | 77 | 2 |





For our Comedy Films, we noticed the general pattern that reviewers actually had a lot more to say about the actors themselves in our highly rated films. There were mentions of Robert De Niro, Lily James, Mila Kunis, Cate

Blanchett, and Anne Hatheway which showcases that reviewers care more about the films cast than other aspects of the film such as character development.

## Recommendations

### For Production Companies

- For comedy films prioritize highly rated actors that people are interested in. Based on our brief research we could see that actors like Mila Kunis and Robert De Niro were generally in high favor for comedy films. They should also anticipate films that are viewed as a "waste of time" either through lack of comedy or even going over the top.

- For drama films they should invest in strong special effects and avoid being cheap with them as well. They should also avoid lack of character development because viewers tend to prefer strong storylines and characters in drama based films.

- For Action films production companies should focus on special effects that looks great specifically on the big screen. This means that they should focus more on the in-person experience due to reviewers highly rating the special effects. However, this also means that action films should avoid having tacky special effects because they become that much more noticeable.