Movisualization

Rachael Fuchs * Alizée Schoen † Claire Yin ‡

Massachusetts Institute of Technology

Figure 1: In the Clouds: Vancouver from Cypress Mountain. Note that the teaser may not be wider than the abstract block.

Abstract

This project visualizes gender discrepancies in the film industry. It includes user-interactive and dynamic, animated visualizations of 600+ IMDb movies. Metadata-rich IMDb movie conversations and ratings datasets found on kaggle contribute to analyses of female vs. male dominance in movie scripts and leading roles. Reading between the lines of the scripts allows for both breadth and depth in understanding the dynamics of gender in Hollywood.

1 Introduction

Movies are an entertainment form that people have been enjoying since the 1880s. However, in the past few years equality, representation, and discrimination have been the topic of discussion. Reading between the lines adds data and information to the discussion to see the breakdown of male and female dialogues, credit's positions, ratings and change over the years in the movies we all know and love. We also have an extended breakdown into our visualizations by genre and by specific movie to explore the changes and effects even further. To analyze these gendered focus insights in the movie industry, we built a variety of visualizations with D3. Not only do we allow users to explore overall comparisons between movies, but we also let users dive deeper into specific movies, conversations, historical dates, and ratings.

*email: rfuchs@mit.edu
†email: amschoen@mit.edu
‡email: yinc@mit.edu

2 Related Work

Existing studies and projects have explored gender disparities in the film industry. One such study analyzes gender disparities in Hollywood using the Bechdel Test. The Bechdel Test was introduced in 1985, and serves as a benchmark to distinguish movies that do contribute to the gender discrepancies from ones that do not. In order to pass the Bechdel test, a movie or story must (1) have at least two women, (2) have women that talk to each other and (3) have women that talk to each other about something other than a man. A film's Bechdel score is determined by the number of constraints met. Using data from the Bechdel Test Movie List, Kaggle, and a ggplot2movies dataset, the author in this study performed analysis and visualization in order to see how many movies pass the Bechdel test, to plot Bechdel scores over time, to find the relationship between IMDB ratings scores and Bechdel scores, etc. Reported findings state that movies that fail the Bechdel test tend to have higher IMDB ratings; an increasing number of movies pass the Bechdel test over time; movies with female directors tend to have higher Bechdel scores [3].

Another project implemented by Anderson and Daniels breaks down film dialogues by gender and age to visualize the gender disparities in Hollywood. Following an analysis of films that fail the Bechdel Test, the authors responded to comments that pointed out the flaws in the Bechdel Test. Their motivation was to use existing data and rhetoric to more objectively view trends that may evidence the white male dominance found in the film industry. Using Disney screenplay dialogue, they visualized the percent of dialoge by actors' age and gender. Their animations and tools allow users to narrow down by genre, high-grossing films, film title, and more. The authors disclaimed that their project was not done to prove a hypothesis, but rather to collect and open source data. From this, the authors' conclusions are simply presenting the statistics. For instance, one

reported statistic is that female characters dominate the dialogue in 22% of films [1].

3 DATASETS

Given that IMDb is the most popular movie website that contains extensive metadata, much of IMDb data exists online through kaggle. We combined and harnessed the powers of these content-rich datasets to create an effective visualization.

3.1 IMDb Conversations Data

Cornell University has a movie dialogue corpus on kaggle that contains raw movie script conversations between characters. The dataset also includes relevant movie metadata such as genres, release year, IMDb rating. Conversational data details which characters are involved, alongside character metadata including gender and position on movie credits. Overall, the dataset contains 220,579 conversational exchanges between 10,292 pairs of movie characters and 9,035 total characters from 617 total movies [4].

3.2 IMDb Ratings Data

Another IMDb dataset contains detailed movie rating data for 85,855 movies. Beyond just one quantitative rating, the dataset qualitatively describes rating details from a demographic perspective. Ratings are broken down by gender and age groups [2].

4 METHODS

Going beyond the Bechdel Test and wanting a more objective, thorough view of gender discrepancies, we used the extensive movie conversational data for a visual analysis of how gender plays a role across conversations in movies. We decided to investigate the relationships between gender, position, and number of lines spoken, incorporating other movie metadata to provide a more holistic overview. The ratings dataset also allowed us to incorporate a demographic (gender, age) perspective into movie ratings.

4.1 Data Filtering

Of the 619 movies within the conversations dataset, 582 of them were also in the ratings dataset. We filtered out the missing 37 movies and worked under the assumptions that they were insignificant movies and that missing this subsection of data would not impact our overall visualization.

Many of the data points for each character have undefined values for gender and the cast member's position in credits. Because of this, we decided to filter out data with these unknown values. Although these characters make up 2/3 of the 9000 characters, we could not gain enough information from this subdata to help our analyses. However, we realized for every movie, the top 10 characters include this gender and positional data. We worked under the assumption that characters without this information are not main characters, thus unimportant for our analysis.

We investigated character positions in credit more by graphing how many movies have data for each position. From this, we saw that the majority of movies do not have character credit position listed past position 8. For instance, 38 characters are ranked 1000-th for cast position. Thus, we only investigated the top 8 characters.

4.2 Design Decisions

4.2.1 Female Line Percentage

One discrepancy we explored is gender dominance within movie scripts. Using the conversation data, we computed the percentage of lines spoken by females and the percentage of lines spoken by males for each movie. We worked under the assumption that the conversations/movies included are samples that are distributed similarly to all conversations/movies.

We utilized the year of each movie to show how this distribution has changed over time. We added in key events in film history to help add context for users, thus make our visualization more effective.

4.2.2 Two Women who Talk to Each Other [3]

Given that each conversation line is associated with a pair of characters who have the interaction, we choose to explore the types of conversations found in each movie. We computed the percentage of female-to-female, male-to-female, and male-to-male conversations. This is a component related to the Bechdel test that we wanted to investigate more in depth.

4.2.3 Female-to-Male Conversation Dominance

For each movie, we used the subset of female-to-male conversations to see how each gender dominates conversations. For each conversation within a movie, we computed the percentage of female lines and the percentage of male lines. This design decision helps us look into the gender discrepancies at a more fine-grained level.

4.2.4 Credits Position

An actor/actresses' position in a movie's credits correlates to their dominance in the film. For instance, position one is the lead of the film. Using this character metadata, we explored the gender distribution across the top 8 positions in credits. This is another facet through which we hoped to point out gender discrepancies. We also computed the average number of lines by gender for each of the top 8 credit's positions as another way to harness the information of this data attribute.

4.2.5 Movie Ratings

We related our movie ratings data to demographics to add additional context to our visualization, and relate gender disparities to popularity across ages.

4.2.6 Breakdown by Genre

One attribute of our data is the genre(s) of the movie, with most movies associated with more than one genre. We choose to breakdown some of our analyses even more by adding tabs for each genre, with the hypothesis that there are larger gender disparities within some genres.

4.2.7 Tabbing

Tabbing breaks down our visualization into four sections: lines, credits position, year and rating. Having this tabbing allows us to present a lot of different visualizations that contribute to our analysis, while organizing a user's attention to focus on specific attributes of our data.

4.2.8 Graph Choices

For the first graph, "Female Character Lines Broken Down by Movies," we choose a dot plot that allows users to interact with each movie individually. The plot shows the distribution of movies by the percentage of lines given to female characters. Once a dot is clicked, more information is shown on the movie. Tooltips provide users with more information on each data point.

We graph the distribution of each conversation to show how the lines are distributed across females and males, for each conversation. Additionally, we show the distribution of conversation types (gender-to-gender attribute). After seeing gender disparities for each movie, and also in the separate ratings tab, users can learn more about the ratings of the movie with relation to age demographics.

In the credits position tab, the graph on gender distribution in top credits positions is a dynamic bar graph to show discrepancies. The distribution of average lines per position in credits is another dynamic bar graph.

The yearly distribution of female lines and male lines graph dynamically changes across the years, allowing users to follow the progression of how this distribution changes over time.

4.2.9 Color Scheme

Defying the social norm of associating blue colors with males and red/pink colors with females, we chose a beige/peach color to represent males and a teal color to represent females.

5 RESULTS

5.1 Female Character Lines Broken Down by Movies

This graph portrays the percentage of female character lines in every movie. This graph can be filtered by genre. When all movies are shown on the graph we can already clearly that the graph is skewed towards left, which signifies most movies has a majority of male lines. Most movies are between 10% and 30%, and there are barely any movie that have more than 60% female lines. This shows a huge gender inequality in movie dialogues. The movie genres that have important trends are Adventure and Romance. In today's society adventure is associated to the male gender, and we can see that almost all adventure movies only have between 10% and 30%. On the other hand, romance movies are often associated with women. When we filter the graph on the romance genre, the data is skewed more towards 40%. So this dialogue graph emphasizes the predictions we could make on the gender inequality in movies.

5.2 Movie Conversations Breakdown

For each movie, this visualization has two components: a pie chart of conversation types, and a bar chart for female-to-male conversations. The pie chart effectively shows gender discrepancies within movie dialogues. One insight from this visualization is that many movies that have higher female line percentage are still dominated by male-to-female conversations. On the other hand, movies with higher male line percentage are dominated by male-to-male conversations. This discrepancy aligns with the ideals of the Bechdel test, and shows that female-to-female interactions are relatively lacking compared to male-to-male interactions.

Then, the second chart breaks down female-to-male conversations. We attempted to show possible gender discrepancies through this visualization. This analysis shows male conversation dominance in some movies. We showed this visualization as an objective analysis that could lead users to see these disparities in line percentage.

5.3 Gender Distribution of Top 8 Credit's Positions

This visualization portrays the percentage of female/male actors at each credit's position. In movies, the higher the credit's position of an actor the higher the pay is (1 being high - 8 being low). The skew of male actors and female actors at each position is prominent with the diverged stacked horizontal bar graph for each genre. The y axis centered at the middle is a 50% mark showing where we would like to see the bar graphs of male and females joining together are sadly not even close a majority of the time and highlights this inequality in movies. In each genre tab, it is clear that males predominantly hold the 1st credit's position over the females. This holds true for every genre and position except 1. The key credit's position to note where the female actors gets closest to the 50% or over is in all genres of the 2nd credit's position. Specifically, the Romance genre is the only position where women are over 50% in the 2nd credit's position because of heterosexual couple leads being the most common for movies. Thus, if there were two main leads in the movie male and female, unfortunately the women are more like to be bumped to the 2nd position or lower when more male leads are in the cast. This is an unfortunate trend to see, but hopefully seeing the facts will enact change.

5.4 Average Lines by Gender for each Credit's Position

This visualization depicts what percentage of lines a male/female character has given their credit's position in the film. No matter what genre we filter this graph on, when the credit position one is a male, they always have more lines than if the credit position one was a female. So even when a woman gets the first credit position, they are still discriminated. The smallest difference in the percent of lines of position is found in comedy movies. In thriller, romance, and drama genres, the difference is still big for a credit's 2nd position, where the male dominates again. Surprisingly, we found that in adventure and action movies, the female dominates in the number of lines for the 2nd position! Finally past the 2nd position, the lines vary between different genres. To conclude, this graph showed us that even when a movie seems to be "equal" (when a woman is in position one in the credits), the inequality in pay and representation is still there.

5.5 Average Yearly Distribution of the Difference of Male and Female Lines

This visualization shows the average percent difference between males and females every 5 years from 1935 to 2010. We added some important events and milestones related to women in the entertainment industry that happened during those years. In the graph, there are two big dips happening in 1940 and 1950. The events marked around this time period helps us understand the reason for these dips. At the beginning of the 20th century the woman in film genre was created for silent movies. During World War 2, as mentioned in the graph, this genre got very popular, which explains why the difference of lines between males and females went up -15%. In the 1950s, sound started being used in movies, and the era of silent movies, and the woman in film genre ended. Furthermore, movies needed more people behind the set, such as engineers, and these jobs were very male dominated. For these reasons, the difference of lines between males and females went back down to -50%. From there, this difference stayed in the -40s %. Gender inequality in movies have been brought to light only recently. In the 1980s, women started to win awards, and this number slowly went up. We can see a steeper upwards trend starting in the 21st century, which shows that advocating for equality has had an impact in the film industry. Today, men still have on average 20% more lines than women, which is a huge progress since the 1930s.

5.6 Movie Rating Breakdown

The overall goal for the ratings graphs was to show that this one average rating that we see on IMDb or Rotten Tomatoes isn't very accurate or applicable to how we as a viewer feel about a movie. Luckily, there is data on all the breakdowns and details of the overall rating for movies on IMDb and we were able to visualize it.

The results from the "Overall Ratings" visualization, already showed us that this average rating just broken down by U.S and Non-U.S users is already different to this average rating. So, viewers can understand a more detailed version of the rating and what might relate to them more whether they are from the US or not. In addition, the number of people voting is very useful to see to comprehend how heavily weighted one person's rating is, and the results are seen further in the "Rating Breakdown" graph.

The "Rating Breakdown" is split into two portions: the bar chart and the gender/age/personal ratings. The bar chart displays the percentage of people for each specific rating in further detail. The results here are pretty expected after viewing the "Overall Ratings" visualization because of how math works with an average. If the average rating is 7.4/10 then the majority of people will be grouped around 6-8. However, it is still useful to see if a 1/2 or 9/10 rating could have skewed the rating lower or higher, respectively. Especially, if there was a lower number of voters, then this graph is useful

to see because 1 vote has a higher weight when influencing the average. The gender/age/ personal ratings portion of this chart allows you to place your rating and see what gender/age you are similar to or not. This allows for a further breakdown to perceive how the overall rating is effected by gender/age as well as seeing what ratings you are similar to. The results of these star graphs of age/gender breakdown ratings aren't too far off from each other but there are some distinctions which was expected. It definitely depends on the movie, but overall the user can fully see who is rating the movies and why a rating is similar to them or not.

One thing to highlight is that even if you identify with a gender/age/race/nationality your ratings could be similar to people with different gender, age, race, and/or nationality. We hope that these results enlighten users to understand what group of people they identify with or are similar to when it comes to ratings. That way, they can get a more accurate rating and prediction of if they will enjoy a movie or not than just an overall average rating clumped together.

6 Discussion

Reading between the lines is meaningful because it draws out the gender inequalities in the movie industry through dialogue, conversation line breakdowns, credit's positions, years, and ratings. Users can easily navigate our visualizations by interacting with buttons, genre tabs, and hovering over data points to reveal more information. We give the users lots of information scent visually and as they move around their cursor. If this isn't enough, we also physically highlight instructions for all charts in the sub-header description.

From our visualization, the audience has learned deeper insights about the gender inequalities in movies than they previously knew. As this is a talked about subject, audiences find it useful to see the data and facts for themselves. The personal further investigation factor has enabled audiences to connect with the topic even more and reveal what is truly behind the curtain of their favorite movies. Audiences leave our visualizations with a new perspective and intuition on the movie industry. It will forever change their outlook on movies and there will be an awareness shift when watching movies (for the better). Thus, the inequalities towards women in the movie industry is brought to the surface and acknowledged by our audiences.

7 Future Work

The dataset we used for this project only includes 617 movies. In the future we would like to find conversation data from more movies. One way to do so would be to web scrape public scripts. With more movies, we will have a more accurate trend in our graphs and users will be able to search more of their favorite movies. Furthermore we would like to obtain movie scripts from the world. From these we could graph out the gender inequality in the film industry in every country. We would be able to compare different countries with each other and find trends with a country's development and gender inequality.

Acknowledgments

We wish to thank the 6.859 staff and our fellow classmates who provided feedback on our visualization prototype. We would like to thank the 6.859 staff, including Professor Arvind Satyanarayan and all of the course's TAs, for providing a rich curriculum and much guidance during the course of this semester, and for answering all of our questions, comments and concerns. Through this class, we were able to build fundamental knowledge of and experience with bringing data to life through visuals and animations. We are grateful for these newly gained skills that will greatly benefit our future academic and professional careers.

References

[1] H. Anderson and M. Daniels. Film dialogue. 2017.

[2] S. Leone. Imdb movies extensive dataset, 2020.

[3] N. Selvaraj. The bechdel test: Analyzing gender disparity in hollywood. June 2020.

[4] C. University. Movie dialog corpus, 2017.