Exploration of Gender Roles in Movies

Rebeka Mukherjee, Archit Rathore, Yash Gangrade

1 Introduction

Gender refers to the socially constructed characteristics of women and men such as norms, roles and interpersonal relationships [1]. The United Nations recognizes gender equality not only as a fundamental human right, but also as a necessary foundation for a peaceful, prosperous and sustainable world [2]. Achieving gender equality is one of UN's goals towards sustainable development.

Media plays an important role in our lives. Though not an accurate representation, but media largely reflects the social constructs of the world we live in. At the same time, media largely influences our daily decisions and way of living. Apart from the general informative form of Media i.e. News, other influential forms of media are Movies, TV series, Documentaries, Songs etc. Specifically, if we talk about movies, they are one of the most crucial source for entertainment in our lives. Nowadays, there have been lot of movements regarding the gender inequality which lead us to the idea of analyzing the movies. In the current scenario, it is important to learn about the portrayal of woman in the Movies because we can easily name quite a few movies where the movies are male predominant in terms of cast and crew and as well as the story-line. We are motivated to learn more about this through the actual numbers. We speculate that the representation of the genders on screen largely reflects our perceptions of the genders in real life.

In this study we explored and analyzed the role of gender in movies. We want to take our results and findings and compare them with the current pillars of gender equality in the world. There are many questions our analysis can answer and we can also learn more about the general trends over the years or equality in different genres etc. For ex, how can media be used efficiently to achieve equality among the genders?, will the gender inequality increase or decrease based on the current data? etc.

The following sections describe our project in detail. Section 2 provides information about the data that we used for the project, how we processed it for the study, and the preliminary exploration done on it. Section 3 focuses on the approaches taken by us to meet the project objective. The findings of the approaches are reported in section 4. Section 5 documents the evaluations of these findings. Section 6 and 7 touch up on the future directions of this project and the conclusions we drew from this project, respectively.

2 Data

The complete data is fetched from this Kaggle dataset: `https://www.kaggle.com/rounakbanik/the-movies-dataset/`. The dataset is downloadable after creating a Kaggle account and comes with a public domain creative commons license.

The dataset contains metadata for 45,000 movies. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The compressed dataset is roughly 230 MBs. The ratings file contains the rating obtained on 45000 movies from 270k users of the website. The possible ratings are with 0.5 discreteness level and ranging from 0 to 5. On the similar note, the credits file comprises of the information about the cast and crew of the movies. Finally, the movies metadata contains information about each movie based on more than 20 features like Revenue, Vote Average etc.

For the text analysis and creating the bag of words model of the scripts, we developed a web scraper to extract the script of the available movies from the Internet Movie Script Database (IMSDB) website: `https://www.imsdb.com/`. We are extracting about 1030 movie (nearly the entire data of IMSDB).

2.1 Data Processing

Though the downloaded data was in CSV format, there were some issues with it like missing values, invalid values etc. Till the data collection report, we performed all the cleaning of the data by removing non-essential columns and the invalid data rows from it. We created objects for each movie by using information from the files 'movies_metadata.csv' and 'credits.csv'. Figure 1 shows all the member variables and methods associated with each object. We used these collection of objects for all preliminary data exploration.

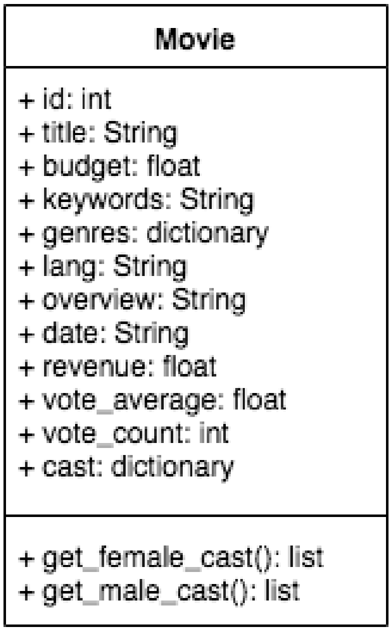

Figure 1: Class diagram for the Movies class

As discussed in the intermediate report, We also wanted to convert the data into a graph that could be further used to perform spectra based analysis methods. For this purpose, we randomly selected 300 English movies whose script was available in the IMSDB website. The reason we selected only English movies is because we speculate that it is the language in which most number of movies are viewed in the United States. English movie scripts were also easily available from the IMSDB website, and easy for us to process and analyze. Although we have scripts from about 1000 movies scraped and stored from IMSDB.

We came up with a list of gender words (Table 1) and created a bag of words vector for each of these movies based on how many times these words appeared in the script for that movie. This table is just a subset of the words we used, we understand that there can be several additions to it. Table 2 shows the corresponding vectors for a subset of 3 movies.

Male Words	**Female Words**
he	she
him	her
his	hers
himself	herself
man	woman
boy	girl
lord	lady
sir	madam
father	mother
grandfather	grandmother
son	daughter
grandson	granddaughter
brother	sister
husband	wife
boyfriend	girlfriend
uncle	aunt
nephew	niece

Table 1: Gender words used to generate bag of words vectors

We found the pairwise similarities between two movies by taking the cosine similarity (1 - cosine distance) between the bag of word vectors for the corresponding movies. Then we represented all the movies as an undirected graph by constructing an adjacency matrix where each movie was a vertex, and the weight of the edges between them was their pairwise similarity. We had to additionally adjust the adjacency matrix by setting the diagonal elements to 0. Table 3 shows the adjacency matrix created from the movies mentioned in Table 2.

Movie Name	BOW Vector
Toy Story	[158, 86, 220, 27, 14, 13, 0, 4, 0, 0, 1, 0, 0, 0, 0, 0, 0, 11, 25, 0, 0, 0, 0, 1, 0, 6, 0, 0, 0, 3, 1, 0, 0, 0]
Braveheart	[411, 184, 494, 17, 45, 15, 41, 13, 60, 0, 32, 0, 17, 16, 0, 12, 7, 145, 187, 1, 6, 10, 12, 4, 1, 9, 0, 7, 0, 0, 9, 0, 0, 0]
Notting Hill	[181, 50, 68, 5, 31, 3, 1, 10, 0, 0, 1, 0, 2, 0, 6, 0, 0, 152, 104, 1, 2, 8, 22, 1, 1, 3, 2, 0, 0, 2, 2, 1, 0, 0]

Table 2: Bag of words vector generated for 3 movies.

	ToyStory	Braveheart	Notting Hill
ToyStory	0	0.975	0.668
Braveheart	0.975	0	0.817
Notting Hill	0.668	0.817	0

Table 3: Adjacency matrix generated for the above 3 movies.

2.2 Data Exploration

To start off in this study, we wanted to learn more about the general relationship between the movies and the gender ratio in it. For inital analysis, we just used the English movies which constitutes the majority of the movies out there. We found out that out of 32937 English movies, only 4457 movies had a cast where the number of female cast members were more than the number of male cast members i.e. just **13.53%**. This preliminary result shows us that majority of the movies are male-centric rather than being gender equal.

Next, to make our results more concrete, we computed the female-to-male ratio for each genre of movie. To do this we simply extracted all the genres of English movies and for each genre, we found the ratio of the number of female cast members to the number of male cast members for every movie in that genre. Then we took the mean of the ratios for every movie in a particular genre. Table 4 shows the results that we found from this step.

Genre	**Female-to-male Ratio**
Action	0.395
Adventure	0.425
Animation	0.457
Comedy	0.675
Crime	0.479
Documentary	0.155
Drama	0.678
Family	0.689
Fantasy	0.607
Foreign	0.521
History	0.324
Horror	0.678
Music	0.6
Mystery	0.652
Romance	0.814
Science Fiction	0.494
TV Movie	0.877
Thriller	0.595
War	0.295
Western	0.319

Table 4: Female-to-male cast ratios for each genre of English movies.

There can be quite a few inferences which we can draw from these small amount of results. It is very surprising to see that ‘Documentary’ movies have the least female-to-male ratio, followed closely by ‘War’ and ‘Western’ movies, which is not so surprising. On the other hand, ‘TV Movies’ have the largest female-to-male ratio, followed by ‘Romance’ and ‘Family’.

Again to extend our results, we computed the female-to-male ratio according to language of the movie. To do this we extracted the languages of all movies and for each language, we found the ratio of the number of female cast members to the number of male cast members for every movie in that language. Then we took the mean of the ratios for every movie in a particular language. While doing this we ignored any language that had less than 5 movies in the dataset. Table 4 shows the results for the top 20 movies with the highest female-to-male ratios.

Vietnamese movies apparently have the largest female-to-male cast ratio which is quite close to gender equality, followed by Tagalog and Korean movies. Surprisingly, all the three languages are from East Asian countries.

Language	Female-to-male Ratio
Vietnamese (vi)	0.994
Tagalog (tl)	0.787
Korean (ko)	0.72
French (fr)	0.682
Latin (la)	0.667
Esperanto (eo)	0.667
Kurdish (ku)	0.667
Spanish (es)	0.653
Abkhazian (ab)	0.648
Japanese (ja)	0.639
Italian (it)	0.623
Hindi (hi)	0.606
Swedish (sv)	0.596
German (de)	0.594
English (en)	0.566
Catalan (ca)	0.546
Portuguese (pt)	0.546
Chinese (zh)	0.54

Table 5: Top 20 languages with the highest female-to-male cast ratios.

3 Methods

Through the initial data exploration and basic analysis of the results we moved on to trying more sophisticated techniques and methods to get meaningful results from the data. Following are some of the approaches we tried to understand the data from different perspectives.

3.1 Approach 1: Regression

Our first approach was to see if we could fit a least squares linear regression model to our data.

First, we built a very simple model to predict the average female-to-male cast ratio of English movies, given the year the movie was released. To do this we found the female-to-male cast ratio for each movie released in a particular year. Then, for each year we found the mean of these ratios.

Using this we trained our model.

Then, we increased our feature set by creating and adding a bag of words vector from the overview text of each movie to the date of release. We used the gender words from Table 1 to create our bag of words. We used this to predict the female-to-male cast ratio for each movie. Here is an example of the features for the movie ‘Toy Story’: [1995, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]. The y value for this movie is 0.333. Thus, having an extended feature set is better because then we can say our regression is reliable.

3.2 Approach 2: Similarity

Our second approach was to compare movies based on similarity. For this, we used the graph mentioned in section `2.1` over 300 movies where each node is a movie and the weight of the edge between two movies is the cosine similarity of the bag of words vector generated for the scripts of the two movies using the gender words mentioned in Table 1. If needed, the cosine similarity can be changed to some other distance metric but for now it’s a good choice.

We used a movie within this subset with the female-to-male ratio of 1 as the seed, and looked at all nodes it was connected to with a weight greater than or equal to 0.95. Then we compared the female-to-male cast ratios of these movies to the average female-to-male cast ratio of English movies (as given in Table 5, in order to validate our results.

3.3 Approach 3: Non Negative Matrix Factorization

Another approach that we have tried here is by applying the Non-negative Matrix Factorization methods to get a vector representation of the movies which we can use later to apply some techniques like regression. First, we are using the *ratings.csv* to fetch all the user id, movie id and rating and then we are finally creating a sparse matrix. Sparse matrix is a good option since there are 45000 movies and storing them directly is not recommended. Once we have the sparse matrix, next we apply the Truncated SVD decomposition (`https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html`). Truncated SVD in a way can be considered as a counterpart of the PCA i.e. Principal Component Analysis for Sparse Matrices. Finally, we are applying the non negative matrix factorization (`https://www.wikiwand.com/en/Non-negative_matrix_factorization`) to this sparse matrix. Here this sparse matrix V is factorized into two matrices (can be more if needed), with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect (Source: wikipedia). After doing all this, we have got a vector representation of the movies which we can further join with the metadata of movies by movieId. We can apply various techniques like Linear Regression, Ridge Regression etc. to this vector representation of movies.

3.4 Approach 4: Clustering —— It is directly taken from Intermediate Report, should we keep it?

We are using different techniques of Clustering of the data we have after the required processing and storage. Before performing clustering, we had to get all the unique Genres, all unique languages from the films and so on. Once, we have that, we just enumerate these unique items in form of a dictionary. We use these enumerations later in forming the numpy array. To create a numpy array, we then iterate over all the movies in the dataset (i.e. about 45000) and then get the fields like revenue, vote average, etc. in the numpy array. Each of these fields is associated with a particular movie and can be used for categorical clustering. Then, we pass this resultant numpy array to the scikit-learn Kmeans function. We are using scikit-learn 0.20.3 version and we are using the in-built Kmeans function. https:// scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. Similarly, we are also using the Spectral Clustering defined in the scikit-learn library. We define the number of clusters as 20 here since we have 20 different genres. Apart from that, we are using the default affinity matrix and not using the precomputed options.

4 Results

Following subsections compiles the results from different approaches we took on to analyze the dataset.

4.1 Approach 1: Regression

Figure 2 shows a plot of the female-to-male cast ratio against the corresponding year using blue points. It also shows the linear regression model that was fit on the data points using a red line. We can see that the ratio increases steadily with each year.

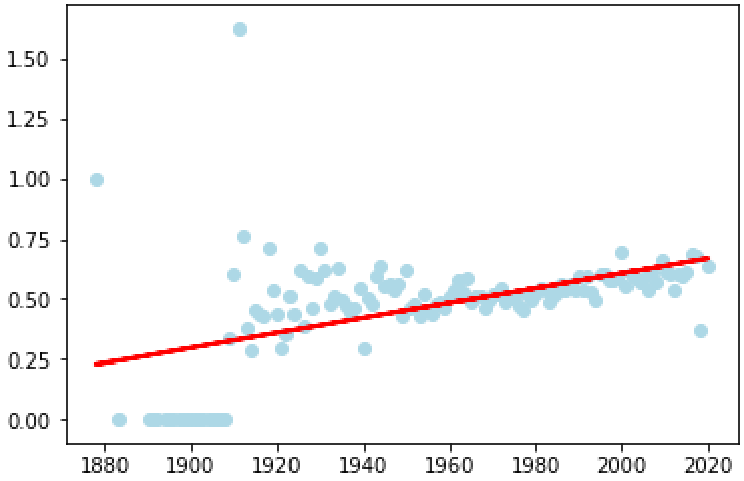

Figure 2: Female-to-male cast ratio vs. movie release date

We would ignore the data about movies before 1920 because movies were not a popular and influential form of media back then. Data about movies were also probably not documented properly back then.

We used our model to predict the female-to-male cast ratios in the next few years. Table 6 shows some of our findings. According to the model, the future looks promising for gender equality.

Year	Female-to-male Ratio
2020	0.669
2030	0.701
2040	0.732
2050	0.763

Table 6: Female-to-male cast ratio predictions in the next few years.

Till this point, we were using the basic features for regression like ratio, year etc. As we discussed in the methods section, to make regression more robust we add extended features like bag of words etc to the model. After training the regression model with the extended features, the following are the results we got:

Coefficients: 0.00196131, -0.00938181, 0.00060612, -0.01883477, -0.06328047, -0.04728943, -0.01352373, -0.10219383, -0.04466916, -0.0184048 , -0.01386901, -0.00150656, 0.1008569 , -0.02415882, 0.08578996, 0.02704775, -0.02956841, 0.09406215, 0.04004039, 0.0693289 , -0.07280574, 0.11247093, 0.11776258, 0.16313531, 0.0632939 , 0.41049857, 0.14284822, 0.1705879 , 0.13669169, 0.22338932, 0.11636392, 0.0834649 , 0.14178295, 0.11000807, 0.07356729

Intercept: -3.3963871812633526

Here we see again that the slope of the linear regression is positive, suggesting that the female-to-male ratio increases with the increase in year.

4.2 Approach 2: Similarity

We found that out of the 300 movies, the movie 'Jane Eyre' had the highest female-to-male cast ratios of 3.75. We then found out all other movies in the graph that had an edge connected to 'Jane Eyre' with a weight of 0.95 or greater. Table 7 lists all these movies and their corresponding female-to-male cast ratios.

Movie	Female-to-male Ratio
My Best Friend's Wedding	1.0
Beloved	2.5
Even Cowgirls Get the Blues	0.875
Scream 2	1.737
Drop Dead Gorgeous	1.714
Scream	0.545
Carrie	1.125
Aliens	0.467
G.I. Jane	0.182
Jane Eyre	0.667
Stepmom	1.667
The Long Kiss Goodnight	0.444
Prom Night	0.6
Heavenly Creatures	1.5
Cruel Intentions	1.571
All About Eve	1.375
10 Things I Hate About You	0.6
Friday the 13th	0.727
Never Been Kissed	0.75
Heathers	0.778
Peggy Sue Got Married	1.125
The Haunting	1.5
Serial Mom	1.625
The Piano	0.5
Twilight	0.5
The Birds	0.875

Table 7: Movies similar to 'Anastasia'.

We know from Table 5 that the average female-to-male cast ratios for English movies is 0.566. We found that 20 movies of the 26 movie **76.9%** more than 95% similar to 'Anastasia', a movie with female-to-male cast ratio of 1, had a female-to-male cast ratio greater than or equal to 0.566.

4.3 Approach 3: Non Negative Matrix Factorization

4.4 Approach 4: Clustering

After getting the clusters by the method we described earlier, we are then computing the dominant label for each of the genres and also computing the frequency distribution of male vs female for each of the genre cluster. Below is a snapshot of a part of the table comprising the result.

Genre	Percentage of Female Cast	Percentage of Male Cast
Comedy	0.333	0.667
Family	0.238	0.762
Adventure	0.323	0.677
Fantasy	0.349	0.651
Crime	0.316	0.684
Thriller	0.25	0.75
Science Fiction	0.375	0.625

5 Evaluation

Due to the exploratory nature of this project, we did not have any metrics to evaluate our results, or any values to compare our results to. However, as mentioned earlier, our main objective was to compare our findings with the current situation of gender equality in the real world.

We found that ...

6 Future Work

Further work on this topic would include applying topic modelling techniques on the movie scripts to gather in-depth information and be able to answer more interesting questions. We can extend some analysis like Bechdel Test [3] to gain information about representation of woman in the films. Further complicated analysis involves scraping the trailers/sneak-peaks of the movies and then applying image processing, machine learning to gain insights about on-screen presence of female cast etc.

7 Conclusion

The issue of gender equality is very relevant in the current times. Mining data from movies and using proper visual techniques to represent the data is an interesting way to explore the topic.

References

[1] World Health Organization. Gender, equity and human rights.

[2] United Nations. Gender equality.

[3] Alison Bechdel. Bechdel test.