Exploration of Gender Roles in Movies

Rebeka Mukherjee, Archit Rathore, Yash Gangrade

Contents

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.

Recently, there has been a lot of discussions on gender equality. This gave us the idea of analyzing movies to study the current state of gender equality. We thought it could be interesting to study the portrayal of women in the movies because we can easily name quite a few movies which 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 project our goal was to predict the female-to-male cast ratio for movies by creating vector representations for the movies and then applying regression models to the vectors and evaluate the performance of different kind of representations. The focus of the project is to efficiently create these vectors for the very large dataset we have while working with the memory and processing limitations of our hardware. The representations we used are: hand-designed features from metadata, bag-of-word features from the plot summary (and scripts for a small subset) and matrix factorization techniques, namely non-negative matrix factorization and truncated SVD.

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 the Kaggle dataset here [3]. The dataset can be downloaded by creating a Kaggle account and comes with a public domain creative commons license.

The dataset contains metadata for 45,000 movies. Each instance contains 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. Similarly, the credits 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 attributes like Revenue, Vote Average etc.

For 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.

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 features 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.

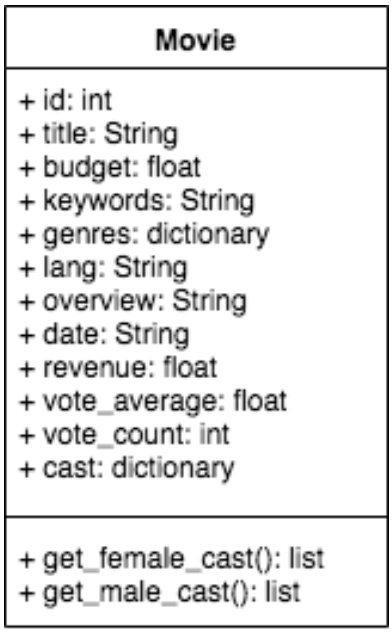

Figure 1: Class diagram for the Movies class

For the plot summary representation of the movie, we scraped the scripts of the movies from IMSDB and created the bag of words vector. We also are creating a vector representation of the movies using Non-negative matrix factorization. To perform this factorization, we had to store the data in form of a sparse matrix with compressed row representation. Here, we are linking 45000 movies with 270000 user ratings and thus sparse matrix where you only store non-zero entries is essential to comply with the memory and processor constraints of our systems.

2.2 Data Exploration

First, we wanted to learn more about the general relationship between the movies and the gender ratio in it. For initial analysis, we only used English movies since majority of movies produced and watched are in English. 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 1 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 1: 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 2 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, followed by Tagalog and Korean movies. Interestingly, 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 2: Top 20 languages with the highest female-to-male cast ratios.

Apart from the filtration, we also tried a second approach 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 4.

We used a movie within this subset with a female-to-male cast 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. A female-to-male cast ratio of 1 suggests that the movie has an equal number of female and male cast members, meaning that there is considerable gender equality in the production of the film. 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 2, in order to validate our results.

We found that out of the 300 movies, one of the movies that had a female-to-male cast ratio of 1 was ‘Anastasia’. We decided to use this as the seed. Then, we found out all other movies in the

graph that had an edge connected to 'Anastasia' with a weight of 0.95 or greater. Table 3 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 3: Movies similar to 'Anastasia'.

We know from Table 2 that the average female-to-male cast ratios for English movies is 0.566. Out of the 26 movies that are 95% or more similar to 'Anastasia', we find 20 movies (**76.9%**) have a female-to-male cast ratio greater than or equal to 0.566.

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 and represent the data from different perspectives.

3.1 Bag of words vector from plot summary

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.

We came up with a list of gender words (Table 4) 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 5 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 4: 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) [4] 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 6 shows the adjacency matrix created from the movies mentioned in Table 5.

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 5: 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 6: Adjacency matrix generated for the above 3 movies.

3.2 Metadata Features

One of the representation of our movies data is coming from the exploration of the metadata of the movies. We are encoding the categorical features as one hot vectors. Some of the features we used are number of female cast, number of male cast, budget, vote average, year of release, popularity, director, genre, keywords etc. We are normalizing the features so they are at the same scales while computing the distances between a pair.

3.3 Non Negative Matrix Factorization

Another approach that we tried here was to apply 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 used *ratings.csv* to fetch the user id, movie id and rating of all the movies, and then we used it to create a sparse matrix. Sparse matrix is a good option since there are 45000 movies and 270000 users. Storing them directly exceeds the memory limits of our system very quickly. Once we had the sparse matrix, we applied the truncated SVD decomposition [5]. Truncated SVD can be considered as a counterpart of the PCA i.e. Principal Component Analysis for sparse matrices. Finally, we applied non-negative matrix factorization [7] to this sparse matrix. Here, the 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 got a vector representation of the movies which we could further join with the metadata of movies by movieId. We applied various techniques like Linear Regression, Ridge Regression etc. to this vector representation of movies to predict the female-to-male cast ratio for a test dataset.

3.4 Regression

Our first approach was to see if we could fit a least squares linear regression [6] 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 estimate if there is a linear trend in the target variable. 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 4 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, 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.

4 Results

Following subsections compiles the results for different approaches of representation of the dataset. A table at the end compiles the results for different method discussed in the previous section.

4.1 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.

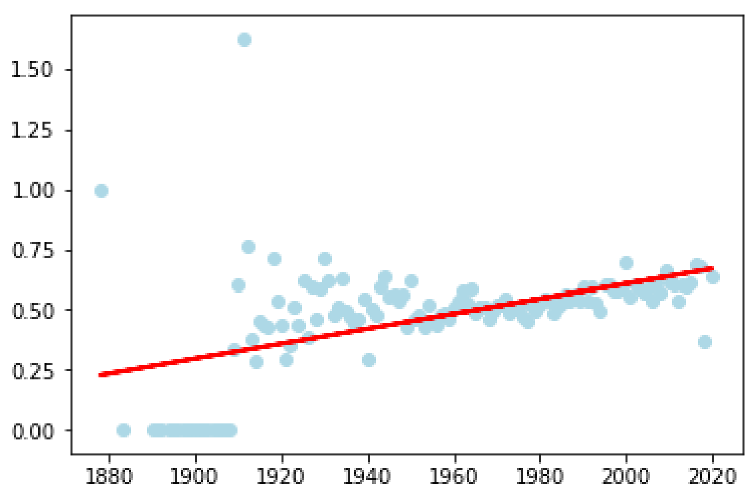

Figure 2: Female-to-male cast ratio vs. movie release date

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 7 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 7: 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 w.r.t year feature is positive, suggesting that the female-to-male ratio increases with the increase in year.

To quantify the 'goodness' of the various features we describe in section 3, we perform regression with each of the feature vectors and predict the female-to-male ratio for each movie and report the RMSE error in table 8. For our experiments we perform 5-fold cross-validation and report the mean RMSE for the experiments. The NMF based feature representations seems to encode the highest amount of information about female-to-male ratio which we use as a proxy for gender equality.

Method	BOW	Metadata features	NMF	Truncated SVD
Linear Regression	0.34	0.27	0.21	0.23
Ridge Regression	0.32	0.21	**0.19**	0.20

Table 8: Root Mean Square Error (RMSE) for linear and ridge regression models for different feature sets. The target variable is the female-to-male cast ratio.

5 Conclusion

We present our findings in this report about potential representation of movies that encode gender bias. We focus on matrix factorization of large sparse matrix as the method of our choice. It allows us to create implicit representations of the data from an existing large corpus. We evaluate the effectiveness of these features in encoding gender equality through proxy of female-to-male cast ratio. The results show that matrix-factorization methods are able to encode this information even though the implicit representation are not derived from gender based attributes. This also highlights the fact gender bias are deeply encoded in our perception and effects even tasks like rating movies. However, we only claim correlation since establishing causality would definitely require a much deeper analysis and control of variables that we haven't considered in our project.

References

[1] Gender, equity and human rights, World Health Organization, https://www.who.int/gender-equity-rights/understanding/gender-definition/en/

[2] Gender Equality, United Nations, https://www.un.org/sustainabledevelopment/gender-equality/

[3] Movies Dataset, https://www.kaggle.com/rounakbanik/the-movies-dataset/

[4] Cosine Similarity in sklearn, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

[5] Truncated SVD, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

[6] Linear Least Squares Regression, https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm

[7] Non-Negative Matrix Factorization, https://www.wikiwand.com/en/Non-negative_matrix_factorization