FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Social Data Science

GENDER INEQUALITY

Does movies contribute to an unequal depiction of genders?

ECTS points: 7,5

Number of characters: 40,072

Date of submission: 30-08-2019

Contents

1 List of contributions

2 Introduction

There has been a persistent debate of gender inequality in several aspects of society, such as education and salary as well as about the representation within different job fields (Evans, 2017). In addition to the persistence of a gender wage gap in most developed economies [1] there is still a tendency of women being largely represented in care-taking jobs, e.g. nursing and kindergarten teachers, whereas males usually are largely represented within management, military professions etc. (Ridgeway, 1997).
This raises the question of where these different perceptions in society come from. According to Whisenant et al. (2002), different kinds of media can be seen as portraying mechanism that create and further deepen gender value portraits in society. This means that the presence and depiction of both sexes in the media, can be of immense importance in order to manifest a change in regards to gender norms.

With this paper, we want to examine how the medium of movies contributes to an unequal depiction of genders. Therefore, we examine several attributes of movies that might influence this gender inequality. Indicators we consider are the ratio of female and male lead actors, as well as directors, and the way it might have changed throughout time. Analyzing movie attributes might lead to insights on how the social perception of women and men is shaped by the movie industry.

In order to analyze this question we have scraped data from www.IMDb.com, where based on the names of the cast, the gender is assigned to each individual. For this, the development over years as well as variation across different genres are analyzed using descriptive statistics. We used, text analysis to investigate the short summaries of the movies and see whether there were a difference in the words or classes of words that were frequently used to describe movies with female or male leads. The reasoning for this part of the analysis was to examine whether there is a diverging description and thus a dissonant portraying of actors with different genders.

Through our descriptive analysis we found, that the share of female leads had increased by 10 pct-points from 1980 until 2019. Nonetheless female leads were still only present as leads in 30 pct of movies in 2019. Further we found that women were mainly present in drama, comedy, romance and horror movies indicating that women might be portrayed in a rather stereotypical fashion in mainstream movies. This result is further underlined by our text analysis of the short summaries of the movies, where we find that movies with female leads are more typically described by words such as "love", "girl" and "home" whereas movies with male leads were described by words such as "war", "man" and "world", thus underlining this notion of movies' tendency to enforce stereotypical behavior.

Further, our analysis found that while men tend to be in movies that make a higher gross revenue the quality of the movies are on average about the same. Finally, we looked into whether female and male directors have different preferences in regards to gender of their lead characters. Here we found that while men chose female leads 26 pct of the time in the 2010s, women chose female leads 58 pct of the time in the 2010s. Thus, increasing the number of female directors might help increase female leads.

[1] data.OECD (2018). Gender wage Gap. Available as: https://data.oecd.org/earnwage/gender-wage-gap.htm. ACcessed 29 august 2019

3 Data collection process

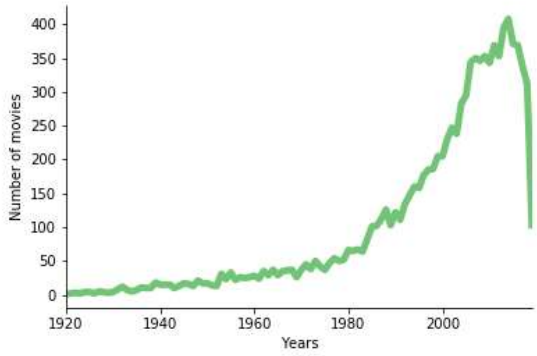
3.1 Collecting data

When searching for a data source, we found that IMDb has an extensive amount of data on movies presented in nicely sorted lists. The website contains information about 6 million movie productions as of May 2019, and has information about movies back to the 1920s. Furthermore, it contains information about 9.8 million movie-related persons. Besides information about release year, actors and directors, IMDb also offers a wide variety of other movie specific data e.g. how the movie has been rated by the IMDb users, the gross revenue, and a short description of each movie.

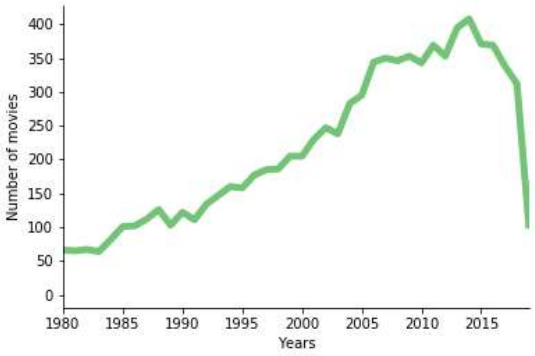
Additionally, the site offers multiple ways of sorting the data, such as by votes, years and ratings. It is therefore possible to filter for preferred information, before starting the scraping process. We chose to sort the movies based on the number of votes received. The reasoning behind this, is that movies which have received multiple votes, have a higher probability, to have been watched by a larger audience as they are most likely more popular (Wasserman et al., 2015). As we wanted to research the amount of female led movies, and its possible impact on society, we preferred the more popular movies, as they have a higher chance of impacting society's view on women.

Figure 1: Number of movies.

(a) Movies 1920 - 2019.



(b) Movies 1980 - 2019



From figure 1a it is clear, that there has been a huge increase in movie releases in the recent decade in comparison to the years 1920-1980. This result aligns with other data movie production history[2].
To have a more unified distribution of movies across the different years, we therefore decided to focus on movies from the years 1980 - 2019, see figure 1b. Thus, there are still particularly fewer movies from the 1980s than in more recent years, but not at the same scale as the 1920s. The abrupt fall at the end of the plot, in 2019 is due to the fact, that the year has not yet ended. Since it is sorted by votes, the newer movies will have fewer votes, since it has only been possible to vote on them for at short period of time.

[2] The Numbers. Movie Production by Year. Available at: https://www.the-numbers.com/movies/production-countries/tab=year Accessed 28 Aug. 2019

.

Unfortunately, while scraping the data we encountered some issues. While more than 100,000 movies appeared from 1980 until 2019, an error in the URL structure made it impossible to extract data beyond page 200. We were therefore not able to scrape more than 200 pages, corresponding to 10,000 movie titles.

In order to overcome this, we considered scraping other movies by using different filters and adding them to the data set however we abandoned this solution due to alignment issues. Adding differently filtered data would dilute the effect of the very reasoning from earlier, where it was argued that having the most voted movies would represent the highest impact on society. Thus, adding movies according to another criteria, such as year of production, might simply include niche movies which are not watched very often and would represent very little impact on society. This approach would therefore weaken the analysis results.

An example of this could be, that having a female lead in a niche movie and a male lead in high-impact movie would represent a 50 pct ratio of male or female representation and one might conclude there is no gender inequality. Then if the female-lead movie is not watched a lot and thus has very little social impact, while the male movie is very popular, the male led movies' effect on society will tend to be higher, and might lead to further deepening in societal gender inequality. Also, there would have been a hit on data quality when it comes to the popularity of the movie. However, it could be interesting for further analysis to add more movies to the data set, once the issue is resolved or data quality can be ensured.

For some of the variables, clarification of their attributes are needed. Firstly the lead role is the actor whose name is presented first in IMDb's presentation of the movie. Secondly multiple genres can be attributed to one movie, thus one movie can have from one to three genres. Finally, the gross revenue is not the worldwide revenue of the movie, but instead the US box office data.

The scraped data was saved as a CSV file on the 23rd of august to ensure robustness of the collected data. This initial data set contains 10,000 movie **Titles** with the respective variables: **Year**, **Genre**, **Rating**, **Gross in M$**, **Director**, **Summary** as well as **Lead** and 3 further "Stars" (**Star 1**, **Star 2**, **Star 3**)

3.2 Collecting names

In order to be able to assign the genders of directors and stars of the respective movies, we chose a methodology of comparing names with gender-classified lexicons. Therefor, we found gender based name lexicons that were used as a basis for assigning gender. In the NLTK package we found two lexicons that could be imported directly into Python. These lexicons contained not only American but also international names, and covered about 7,000 different names. With this approach we ensured the ability to investigate possible gender inequality not only in Hollywood movies, but in movies from all over the world.

After further inspection it was noticed that some of the most frequently occurring leads had very unique names e.g. Charlize Gwynneth, which were not represented in the obtained lexicons. To avoid gender-classification errors as well as observations where matching would not be possible, additional lists of the names of the top actors (681 names), actresses (606 names) and directors (100 male and female names) on IMDb were scraped.

This newly scraped data contained both the first- and surname of the most popular actors and directors.

These lists would therefore make it possible to identify gender based on more uncommon names. Additional names were then added to the previously obtained male or female name lexicons as well as saved as their own string. In the name lexicons in the NLTK package multiple names appeared both in the lexicon of male names as well as the one for female names, making a unique classification challenging. We elaborate on our solution to this challenge in section 3.3.

3.3 Log analysis

To get a greater understanding of our data collection, we have made an analysis of the log we got from the scraping process. The log analyzed was created from a re-scraping of data on the 29th of August, while the data set used in the analysis is from 23rd of August. This has resulted in minor deviance, when the sorting process was applied, but no changes in the size of the primo data set, that still contained 10,000 movies, scraped from 200 IMDb pages. No errors were found after scraping our data. When scraping our data we wrote the code such that we used the same URL, and changed only the number of the page we scrape, thus the URL length stays constant at 216.

In addition to the movies we also scraped the names of multiple actors and directors, but as the data scraped was in a much smaller size, the analysis is divided into two: first an analysis of the log for the movie scraping, then an analysis of the scraping of the names.

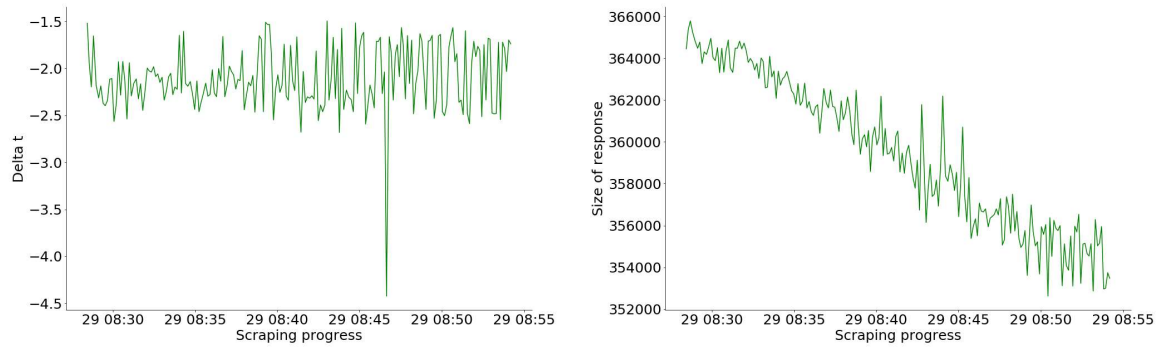
3.3.1 Movie log

From figure 2a we see that for most of the scrapes, the speed varies around a set mean at of approximately -2 delta t. One major deviance occur just after 08:45 where the delta t value takes a major dip down to -4.5, and then directly adjust back to normal. This could be the result of a larger response size, but as seen in figure 2b, there is no considerable outlier at the same level. To check if this could have been because of a data issue, we checked the relevant page data, and found no irregularities compared to the rest of the data. This is not that surprising considering there was no outlier in size, as stated above.

Figure 2: Log of movie scraping

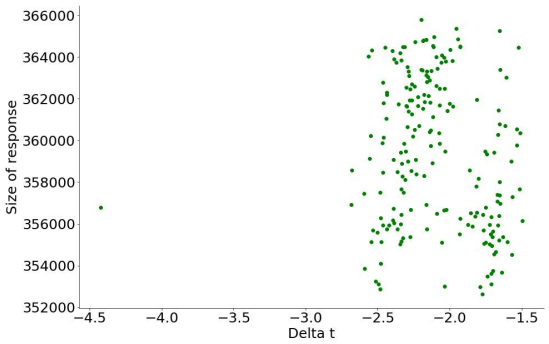
(a) Speed

(b) Size



In figure 2b we see that the size of response is a declining trend, with more or less even variations. The declining trend can be explained by the way we sorted the data as the number of votes i declining the size of the page will be decreasing per default. The variance around the trend is most likely a result of changes in the length of summaries, number of directors etc. varying from page to page.

Figure 3: Speed, Size scatter-plot



In figure 3 we see that most of the scrapes are located in the right side of the graph, with deltas around -2 delta t and varying size, but all within the interval of 350,000 and 370,000. The only outlier from this was the observation from before, which size is in the same general area, but delta t is particularly different. This could have been a result of a drop in internet speed or other external factors as the data from the scrape conformed with the rest of the scrapes.

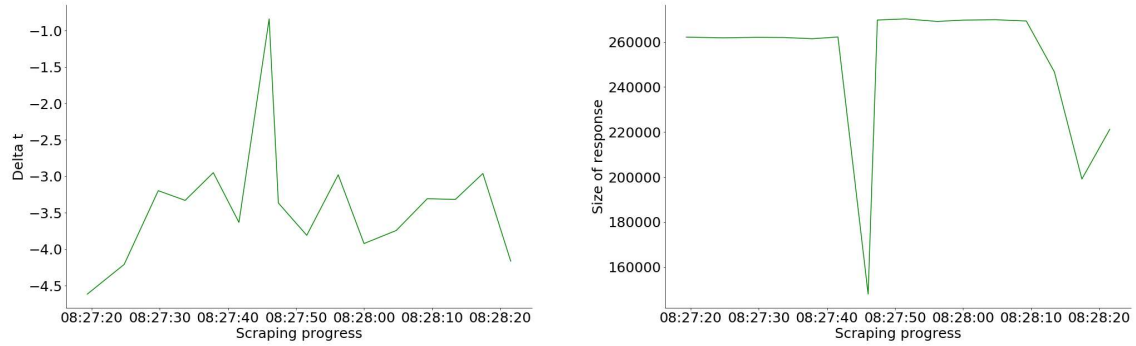
3.3.2 Log of name scraping

When collection the names 16 scrapes were made. This is a very low number, and therefore the log data might not be that informative. Seven scrapes were made for each of the names of the female and male actors, while one scrape was conducted for the female and male directors respectively.

Figure 4: Name scraping

(a) Speed

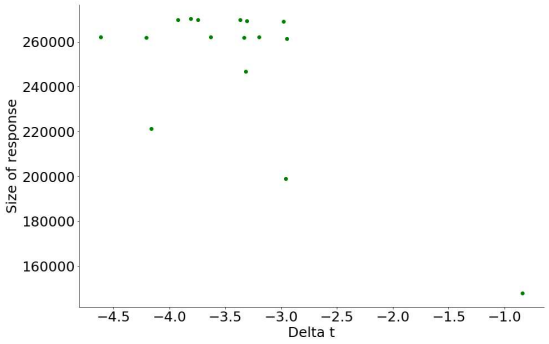
(b) Size



From figure 4a we see that the speed stays around -3.5 Delta t for an outlier at around 08:27:40-50. With

exception of this, the speed is constant. If we compare to figure 4b we also see a remarkable drop in size around the same time. We have found that this is the seventh scrape, and which is the last page for actresses names. This explains the drop in size, as the other pages held the data for 100 actresses, but the seventh, only held six names. Except from this, the changes could be explained by the difference of page layout from actor, actresses and director names.

Figure 5: Speed, size scatter-plot



In figure 5 we see that they gather around the same size, in the top left corner except for some few outliers which are the director pages and the last page of the male and female actor names, which vary in size from the rest.

4 Data cleaning & reparation

Before being able to analyze the scraped data, the data had to be cleaned and prepared. This data cleaning process can be divided into "Data types", "Splitting column values", "Assigning gender" and "Final data set".

4.1 Data types

The data in the **Gross in M$** variable included string characters such as 'M' and '$', which have been removed. Moreover, we noticed that some of the data, such as **Gross in M$** and **ratings**, had been scraped and turned into strings rather than integers or floats. In order to be able to analyze the numerical values in the data set, these values had to be transformed into floats or integers. After this transformation, descriptive statistics such as mean, max and min, could be applied and yield a meaningful results. This gave us a brief and fast overview of the data.

4.2 Splitting column values

4.2.1 Genre

When inspecting the data, we saw, that each movie was assigned up to three genres. For movies with three genres, all three genres were stated alphabetically as a list in one column. We therefore split the column into "Genre 1", "Genre 2" and "Genre 3", to ensure that each column contained only one genre. If we were to pick

only the first genre in the column we would obtain biased results due to the alphabetical ordering. The genres with letters from the first part of of the alphabet would be over-represented compared to those starting with letters from the last part. By splitting the columns and then analyzing all of them separately rather than only the first genre, an unbiased statistical analysis was ensured.
After restructuring the genre columns, we saw that the most frequent genres are drama, comedy and action. This is not a perfect measure, as some genres like comedy often occur together with other genres, e.g. romantic-comedies.

4.2.2 Director

To measure the amount of movies directed by each director we encountered the issue of movies having multiple directors. Sometimes the multiple directors would be a pair of directors (e.g. Cohen brothers or the Wachowski siblings), who have made all of their movies together. Other times directors who have been productive on their own also made co-directorial movies. To account for this we split the names, and only kept the first listed director. When assigning gender to director we used the original data, with multiple directors in each cell. We chose this approach in order to circumvent the possibility of multi-gendered productions being assigned the gender of the first name appearing. This way multi-gendered productions with male and female directors was assigned a neutral gender.

4.3 Assigning gender

The respective directors and leads are assigned their gender by matching the first names to the female and male name lexicons described in section 2.2. New columns for each gender category were created (male, female, neutral) in the process of assigning gender. Since some of the names appeared in both the lexicon for females and males, we manually assigned some of them to either female or male, based on which gender the name is most often associated with. Names such as Clair, Ashley and Judith were for example categorized as female names, whereas Justin, Shawn and Daniel were categorized as male names. In case the names were too ambiguous for us to label them, the names were assigned 'neutral'. Examples of this could be 'Alex' and 'Kim'. For a list of the manually assigned names, see the appendix. We are aware that this approach is not perfect, and many names were sorted as neutral to be safe. Here, a machine learning approach could be undertaken to classify the selected names by gender, but as the amount of neutral name were quite low, manually assigning them were feasible and we would expect an equally good result.
After adjusting the lexicons in order to overcome the neutrality we apply a tokenizer-function, to split the actors names into single strings. This way we could match them with the respective lexicons. When the male lexicon was applied, each string in the name, that was matched with the lexicon was counted in a male count column and vice versa for females. The actor/directors with a count above zero would then be assigned that gender. This resulted in some people being counted as both male and female. These actors/directors were also counted in the neutral column. Actors with no match in either lexicon were assigned none. After this initial assignment of gender, we applied the list of actor/director full names, that was scraped from IMDb. In this case we did not use the tokenizer, as we wanted a match on the full name. If an actor/director matched

in this case, it overruled the counts made with the first name lexicons. Lastly all movies with a neutral or non-assigned gender were removed.

4.4 Final Data Set

The cleaning process eventually resulted in a final data set containing 8,264 **observations**, when movies with non-assigned or neutral gendered leads were removed. This was further reduced to 6,921 **observations** when movies directed by non-assigned and neutral gendered directors were dropped. Both data sets contain 13 **variables**, including: **Title**, **Year**, **Genre 1**, **Genre 2**, **Genre 3**, **Rating**, **Gross in M$**, **Director**, **Summary**, "**Lead**", **d_Female** (director-dummy, 1=Female , 0=Male), **Female** (actor-dummy, 1=Female, 0=Male) and **d_Drop**, (dummy for dropping non-gendered directors, 1=gendered, 0=non-gendered).

5 Descriptive statistics & visualizations

5.1 Brief literature overview

The influence of the medias' depiction of genders has been addressed and analyzed multiple times in different parts of the social sciences (Signorielli, 1990). In 2014, Galdi et al. (2014) found that men who were exposed to objectifying TV i.e. where women were portrayed as sexual objects, were more likely to manifest gender-harassing behavior. In 1990, Peterson and Lach (1990) investigated the effect of children's books on gender understanding and found that there was an effect on both self-concept, potential for achievement and perception of others. Further they found, that by the age of four, children were already aware that the *"[...]feminine role is housekeeping, while the masculine role is wage-earning"* (Peterson and Lach, 1990, p. 3). Although times have changed since the 90s this influence of specific depictions of male and female roles on women's choice seems to persist. In a study from 2010, Rudman and Phelan (2010) showed that women primed with traditional gender roles had a reduced interest in traditional male occupation, thereby enforcing the existing separation of the sexes in the labor market.

While it is easy to find literature in support of the negative effects of stereotypes it is possible to imagine that an empowering depiction of women could pull the tendency in the opposite direction. Thus, generate role models for girls and present a more equal positioning of males and females. In turn this could influence the development of gender norms and equality. In addition to this the movie industry, like many other industries, also face the problem of missing out on talent. Assuming that the innate abilities of acting is the same across genders, talented females might be forgone by less talented males, due to the current gender norms. Finally, a re-enforcement of gender specific stereotypes as well as an acceptance of current gender norms might have negative effects on the equality between men and women in society.

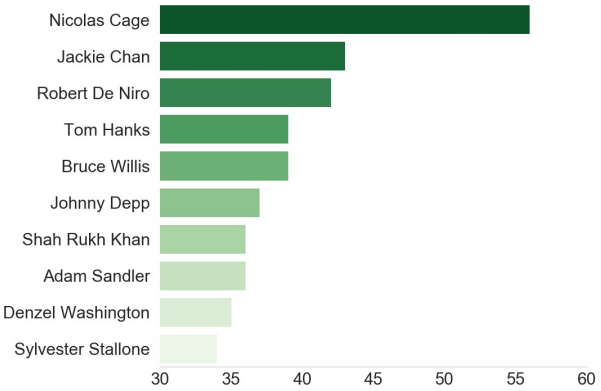
5.2 Men and women as leads

In a time where both children, adolescents and adults are confronted with multiple media every day, it is hard to determine what we are most influenced by. However, because movies have the possibility of creating

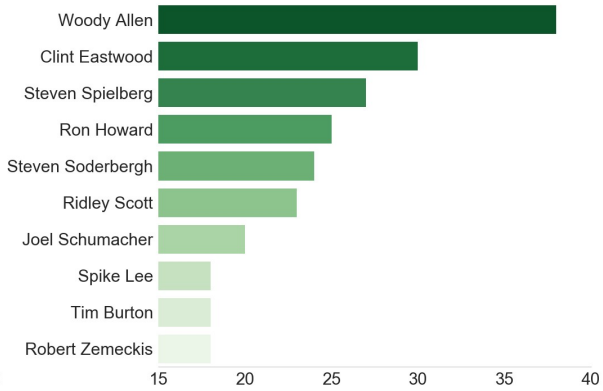
a convincing universe and often demand more of our attention, they have a big impact on us. We therefore argue that this is an important place to look for gender differences and stereotypes. Further the idolization of movie stars might enforce specific world views and norms, which is why we argue that diversity in these areas is important. The data we have scraped from IMDb can give us some understanding of how women have been presented until today in some of the most popular movies.

Figure 6: Top ten actors and directors

(a) Top ten actors 1980 - 2019.



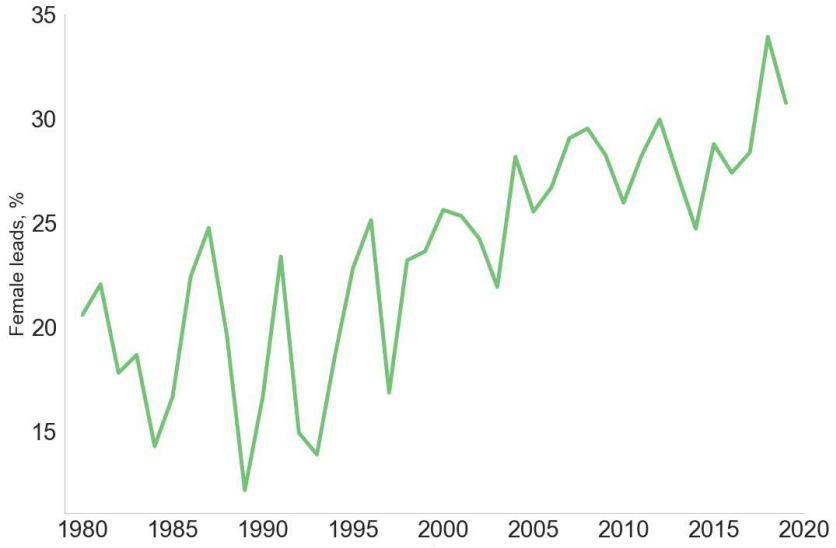
(b) Top ten directors 1980 - 2019



First off, we consider the top ten leads and top ten directors. Here, it is important to note, that top ten is an expression for how often an actor or director have occurred in our data set. For example, the fact that Nicolas Cage is taking the lead in figure 6a, because he has the lead role in more than 40 of the movies in our data set.

Further, taking into account that around 3/4 of the data sample are men the result below might not be too surprising, however it is curious that no women made it to either of the lists. This result should further be seen in relation to the fact, that the share of female leads and directors have been much lower than today (and is still not equal), meaning that the male leads and directors are both dominating in numbers within each year, but also across years. Thus, by illustrating this point through a simple frequency analysis we might exclude females per default. Nonetheless, if we consider only the years from 2010 and onward, we still find no females on either of the top ten list.

Figure 7: Percentage of women in the IMDd movie list



In figure 7 we took a closer look at the development of female leads from 1980 until 2019. From this figure we see that there have been a steady increase in the share of female leads across the period, with a total increase of approximately 10 pct-points. Despite this positive development, the share of women is still well below 40-60 pct From this descriptive analysis we find that despite the positive development, women are still underrepresented as leads in movies. The high variation in the beginning of the graph, can be explained by the lower amount of movies produced per year, as shown in figure 1a and figure 1b, resulting in very unstable ratio-levels, that smooths out as movie production increases over time.

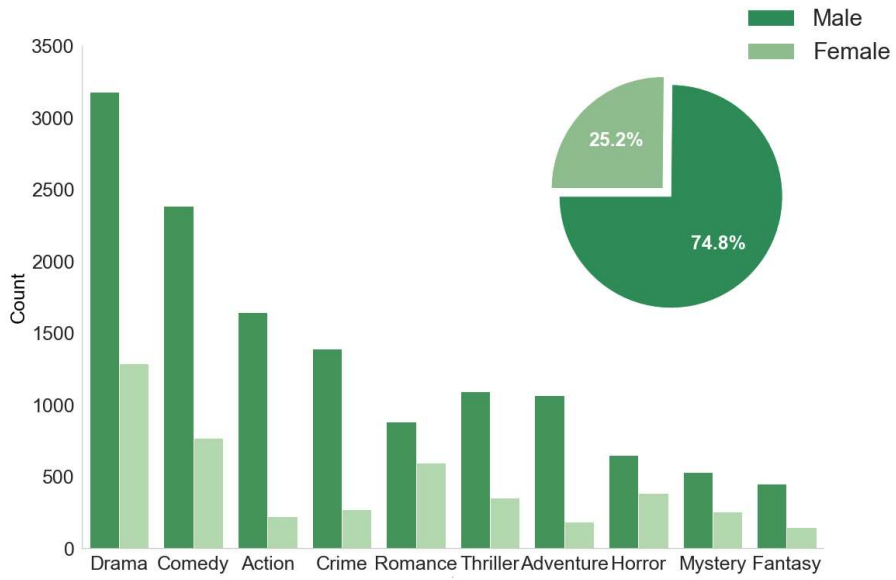
5.3 How women are portrayed

Taking our summary of the literature into account, the positive impact of the development of the female lead shares depend on how women are depicted in the movies. A too stereotypical or too controversial depiction of the female role and characteristics, might actually do more harm than good compared to having no female depiction at all (Rudman and Phelan, 2010). In this section we therefore look into the differences between male and female leads. We do this by locating what genres they seem to dominate, whether there are any noticeable differences in how the movies are described in their respective short summaries and differences in rating of the movies and the gross revenues.

5.3.1 Genre

Because a lot of the movies in our sample fall within multiple genres our approach was to count how many times a specific genre appears with a male or female lead. The result of this is illustrated in figure 8.

Figure 8: Distribution of men and women in the genres



Because this is a frequency analysis it is important to take into account that our sample is not balanced in terms of males and females. When interpreting the results of figure 8 it is therefore useful to keep the gender distribution (illustrated in the pie chart) in mind. From figure 8 we see that females are mostly casted for drama-, comedy-, romance- and horror movies, with romance being the most common genre when compared to men (here there is only 32,2 pct difference between male and female leads, compared to action where there is 86,3 pct difference between male and female leads). The fact that females are not dominating any genres is most likely explained by their under-representation in the sample (which in turn is explained by the lack of movies with female leads through time). However, from this result, we get the impression that women seem to be portrayed in more 'soft' genres than men, who are highly over-represented in the more 'hard-core' genres like action and crime. This might enforce the conventional stereotypes, that can have a tendency to limit both men and women in their ambitions and aspirations of how to live their lives. The only genre which is not considered as 'soft' that women are more represented in compared to the general distribution, is horror. In this genre the lead role is often portrayed as dumb, weak, naive or a victim, not as an empowering role.

Genre on its own, is not a satisfying measure of this tendency has we have to draw conclusions based on heavy generalization. Further, as the story line and universe within each genre can vary, we cannot say, whether the fact that the genre is in a softer category, it is enforcing stereotypes. Therefore, it is interesting to look further into how each movie is described.

5.3.2 Summary analysis

In this part of the analysis we investigated the short summaries of the respective movies. Our goal was to uncover whether movies with female or male lead actors deviate in terms of wording used in these summaries. Different classes of words used could hence indicate how the movies portray the female and male role in

society.

Here we decided to follow an approach which focused on counting the frequencies in which words occur in the short summary for each movie. These counts were conducted for movies with male lead and female lead separately in order to distinguish the results. We tokenized the "short summary" strings, thereafter, we divided the data by gender and investigated the 30 most frequently occurring words. However, some of the tokenized strings turned out to be punctuation signs, written numbers, pronouns or similar non-interpretable strings, lists including these strings were created and a matching approach was used in order to drop the counts respective string.

Figure 9: Wordcloud based on short summary

(a) Male Movies

(b) Female Movies

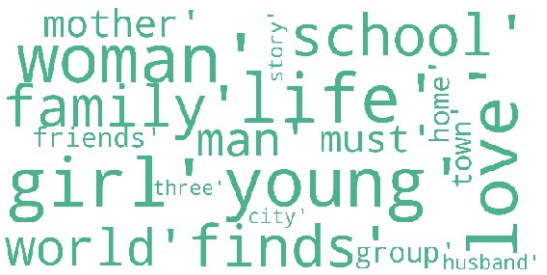


Figure 9 displays the results of the above mentioned text analysis separated by gender, and used two different dimensions to represent the included information. The words depicted as well as the different font sizes deliver insights about the usage of the words in the short summaries.

First, the visualization of the different words used shows that regardless of the gender of the main lead, most movies are described with words like "life", "family", "world", "young", "love". Since these words are not very specific and hence describe characteristics that occur in a wide variety of genres as well as movies in general, the highly frequent appearance of these words are not surprising. This could support earlier discoveries of our analysis, were we found that movies belonging to the genres "drama" and "comedy" are the most frequent within our sample. It is worth noting that there are some specific words that only appear in male or female led movies without having an equivalent present in the other. For example the words "war" and "old" appear more frequently for males leads whereas "girl", "home" and "town" was seen more frequently for females. This could indicate, that by being represented in war movies there is created a perception of men being strong, heroic and mature whereas the appearance of "home", "town" and "city" within female led movies might suggest that they are portrayed as care-taking and being at home, responsible for the household. Moreover, the word "girl" is mentioned to describe movie story lines about females, while the equivalent "boy" is not in the top 20 of males. This additionally might lead to the conclusion that females are categorized as more childish and thereby indirectly more helpless and naive.

Secondly, the fact that the fontsize being representative of the ranking of frequencies also show insights that might support some of the conclusions made in the previous paragraph. The frequency in which the words "girl", "school" and "love" occur in female led movies, supports the stereotypical perception of women in

society. Furthermore, the higher ranking of "love" for female led movies in comparison to male led movies, also aligns with earlier analysis, saying that female lead actors are closest to being represented equally with male actors in the genre of romance.

5.3.3 Gross revenue and ratings

In order to find indicators of why women are not being casted as often as leads as men we have looked into whether there might be a difference in the quality of the movies and in the gross earnings. From table 1 we see that male leads tend to be in movies with a higher average **Gross (in M$)** while the Imdb user ratings seems to be somewhat similar. Again our data comes with limitations and as the variable **Gross (in M$)** only accounts for the earning in the United States this result could be a matter of women being in movies that tend to make money outside the United States.

Table 1

	Gross (in M$.)	Rating
Male leads	44.46	6.48
Female leads	30.66	6.32

However, putting this limitation aside there might be a re-enforcing effect hidden in this result. When the movies with male leads tend to have a higher **Gross (in M$)**, there might be a tendency to choose men instead of women in order to cash out the higher expected revenue, disregarding the direction of the actual correlation. As a result, women might be held back by the assumption/expectation of higher **Gross (in M$)** despite the fact that the quality of the movies are approximately the same. In this way women are less likely to be chosen for the 'big money making' production resulting in re-enforcement of the existing tendency.

As with the measure of **Gross (in M$)**, the **Rating** feature also has limitations, as it is a subjective measure, that might fluctuate a lot. Taking into account that the sample of female leads is considerably smaller than the male leads, the women's measure might be difficult to compare to the men's as it might be more volatile.

5.4 Female and male directors

For our final section we wanted to investigate if there was any positive relationship between the gender of the director and the gender of the lead. This relationship is interesting as the director have a considerable influence on the cast of the movie. Despite the fact that the directors can be constrained by the screen play, they still have an indirect influence in choosing what movies to direct.

Looking into this question we found that the under-representation of female directors is even more apparent with zero female directors in 1980 compared to 63 men. In figure 8 we see that there has been a positive development in female directors, however female directors still only represent a share of 7,5 pct in 2018. In addition to the problem of missing out on talent, female directors might also have a larger interest in telling stories about women, and thus might have greater tendency to choose movies with female leads. This result is also in line with the literature on gender diversity in leading positions in the labor market, where multiple

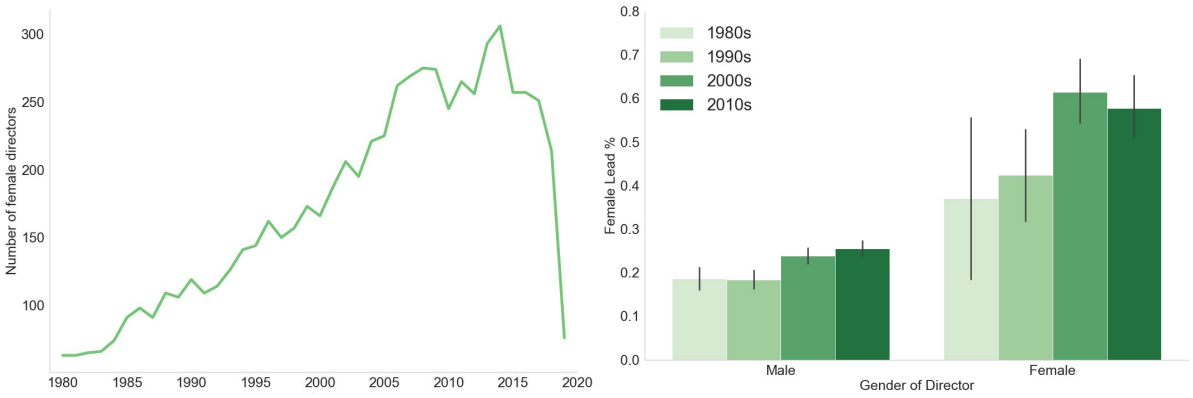
explanations have been proposed in order to explain the lack of women in leading positions in general (Eagly and Carli, 2012).

From figure 10a we see that female directors in our sample choose female leads to a much larger extend than the male directors, confirming our expectations above. Even though there has been a tendency towards more females in both groups of directors, the two groups are still polarized in the 2010s with female directors producing movies with female leads 58 pct of the time while male directors produce movies with female leads 26 pct of the time.

Figure 10: Directors

(a) Number of female directors

(b) Share of female leads



This result indicates that a way of ensuring more gender diversity in the leads of movies could be educating or encouraging more female directors.

Further this result also contain network effects. It is possible to imagine that in addition to female directors being more interested in the female perspective, they might also have more females in their network making it more convenient for them to choose female leads, than their male counterpart. Finally, for further analysis it could have been interesting to see whether female directed movies are in fact less prone to encourage female stereotypes. With a larger sample, this could have been done either in the same fashion as we did the summary analysis or by using a more advanced textual analysis method such as e.g. bag of words. However, due to the limited number of female directors this division was not feasible with our data set.

6 Discussion

Working with data always comes with limitations. In this sections we discuss what consequences the data limitation in our sample, might have on our results.

One of the main issues with our data is the possible selection bias that might occur, due to the fact, that we obtained our sample based on the most voted movies on IMDb. We did this in order to ensure, that we selected the ones with the largest potential social impact. This resulted in our sample being skewed towards

the right as more recent movies have obtained more votes. Even after narrowing down the year interval, we still saw this tendency, since there has been a huge increase in movie releases since 2005, compared to previous years. From figure 7 and figure 10a we see, that the amount of female leads as well as female directors increases through time. Since the presence of female leads and directors probably would be very small in movies from earlier years, we do not believe, that a higher proportion of older movies would have changed our results considerably. However, it might be that the increase in movies is what is driving the increase in female leads and directors.

While collecting data, we were also interested in the gender of the rest of the cast. We therefore collected three more actor/actresses per movie and added these to our data set. It could have been interesting to analyze whether the female led movies generally had more female characters, or if the lead was the only one. Another topic of interest was the difference in gender diversity in the cast for male and female directed movies. We were unable to go through with these analyses due to the time constraint.

Another aspect to consider in our analysis is the skewed distribution between males and females. The frequency of males in comparison to females in the data set can be seen in the pie chart in figure 8, where it follows that 25,2 pct of the samples are female while 74,8 pct are male. This unbalanced property of the data might impact the precision of our results unequally, making it difficult to do a convincing comparison between men and women.

In regards to the text analysis conducted, it has to be kept in mind that these underlying short summaries only represent a very brief storyline and therefore only offer limited analysis potential. It therefore might be interesting to collect further data on the respective movies and investigate the long summaries in regards to potential gender inequality concerns. In order to access these longer summaries, we would have had to conduct and additional very time-consuming scraping process for each individual movie.

A final consideration that should be mentioned about the conducted analysis is the limitations attached to an analysis based solely on descriptive statistics. By using an inference based machine learning approach we could have supported the observed results by further diving into a detailed analysis. Thus, it could be interesting to use regression analysis in order to predict targets like, **Gross (in M$)** as well as **Rating** of a specific movie. This analysis could have helped us in order to get a more detailed understanding of which features are actually significant in regards to predicting these targets. This would further enable us to determine each separate effect of a feature towards the target including. By doing this, we would not just have gotten a generalized knowledge of the data, but more in-depth insights. In reference to 4.3.3, we could then reason which features lead to the observed higher means for **Gross (in M$)** and **Rating** for male leads. This approach could supported us in regards to our research question, because it could have pointed us towards insights of why there are so many more male directors and actors than there are female. Given the fact that this would require additional features which are assumed to be more meaningful in terms of effect on the targets, than the ones included in our dataset, this approach could, however, not be pursued in this specific analysis.

7 Conclusion

Throughout this paper we have investigated how and to what extend female leads appear in popular movies scraped from the website IMDb.com. After having collected the data we, assigned gender based on each actor's, actress's or director's name. The names were assigned through two processes. First we assigned the gender based on first names by tokenizing and then matching the names with a NLTK lexicon. For the names that were then not identified or were represented in both the male and female name lexicons we assigned gender by the full name of the actor, actress or director. The full names were scraped from IMDb lists of most popular actors, actresses and directors. By following this approach we ensured that the individual who were at the end discarded, were not any of the most popular ones. We focused our attention on the lead characters as we argue that these have the largest expected impact on society.

After having cleaned our data set we conducted a descriptive analysis of the depiction of women in the movies in our sample. Here we found that despite a steady increase in female lead shares from 1980 until 2019 (10 pct-points) there is still only 30 pct female leads in 2019. Further we found that women tend to be more present in genres such as drama, comedy, romance and horror which we argue strengthens an existing stereotype, that have been proven to have negative impact on the lives and aspirations of men and women. This result is further underlined by our text analysis of the short summaries describing each movie. Here we see tendency that women are portrayed as more childish and soft, compared to men who, too a larger extent, is portrayed as strong and heroic. Finally we looked into whether there was a relationship between the gender of the director of the movie and the gender of the lead. Here we found that while male directors choose female leads 26 pct of the time, female directors choose female leads 58 pct of the time. Thus, there seems to be a strong relationship between these two outcomes. Additionally, it also emerges from this investigation that the problem of uneven representation of the gender is even more apparent in the context of directors where only 7,5 pct of directors are female in 2019.

From our descriptive analysis we thereby conclude that while the female share of leads in the movie industry has increased, there still seems to be a persistent unequal depiction of genders.

References

Eagly, Alice H and Linda L Carli (2012). "Women and the labyrinth of leadership". In: *Contemporary issues in leadership*, pp. 147–162.

Evans, Mary (2017). *The persistence of gender inequality*. John Wiley & Sons.

Galdi, Silvia, Anne Maass, and Mara Cadinu (2014). "Objectifying media: Their effect on gender role norms and sexual harassment of women". In: *Psychology of Women Quarterly* 38.3, pp. 398–413.

Peterson, Sharyl Bender and Mary Alyce Lach (1990). "Gender stereotypes in children's books: Their prevalence and influence on cognitive and affective development". In: *Gender and education* 2.2, pp. 185–197.

Ridgeway, Cecilia L (1997). "Interaction and the conservation of gender inequality: Considering employment". In: *American Sociological Review*, pp. 218–235.

Rudman, Laurie A and Julie E Phelan (2010). "The effect of priming gender roles on women's implicit gender beliefs and career aspirations". In: *Social psychology*.

Signorielli, Nancy (1990). "Children, television, and gender roles: Messages and impact". In: *Journal of adolescent health care* 11.1, pp. 50–58.

Wasserman, Max et al. (2015). "Correlations between user voting data, budget, and box office for films in the internet movie database". In: *Journal of the Association for Information Science and Technology* 66.4, pp. 858–868.

Whisenant, Warren A, Paul M Pedersen, and Bill L Obenour (2002). "Success and gender: Determining the rate of advancement for intercollegiate athletic directors". In: *Sex Roles* 47.9-10, pp. 485–491.

Appendix

Appendix A

Female	Male	Neutral	Female	Male	Neutral	Female	Male	Neutral
Abbey	Adrian	Addie	Jean	George	Frankie	Sibyl	Theo	Mead
Abbie	Adrien	Alex	Juanita	Geri	Franky	Stacy	Tim	Meade
Abby	Ajay	Alfie	Judith	Germaine	Gabriel	Sunny	Timmie	Mel
Alexis	Ali	Alix	Judy	Gerri	Gabriell	Sydney	Timmy	Merle
Allie	Andie	Allyn	Julie	Gerry	Gail	Tabbie	Tobe	Merrill
Angel	Andy	Andrea	Kelley	Glen	Gale	Tabby	Tobie	Merry
Angie	Austin	Barrie	Kelsey	Glenn	Georgie	Tallie	Toby	Michel
Ariel	Bennie	Barry	Kerry	Gus	Gill	Tally	Tommie	Michele
Ashley	Benny	Beau	Lauren	Hannibal	Harley	Tammie	Tommy	Morgan
Aubrey	Bernie	Billie	Laurie	Heath	Hazel	Tammy	Tony	Morlee
Augustine	Bert	Billy	Lindsay	Henrie	Ikey	Terri	Van	Muffin
Averil	Bertie	Blake	Lindsey	Ike	Jackie	Trace	Vin	Nat
Blair	Bill	Bobbie	Lindy	Jan	Jaime	Tracey	Vinny	Noel
Brandy	Bo	Bobby	Lorrie	Jermaine	Jamie	Tracie	Wallie	Ollie
Brook	Brett	Britt	Luce	Jerrie	Jere	Tracy	Wallis	Pat
Brooke	Chad	Bryn	Lyn	Jerry	Jess	Vinnie	Wally	Pen
Brooks	Chris	Cal	Lynn	Joey	Jesse	Whitney	Willi	Quinn
Carey	Christian	Cam	Maddie	Jordan	Jessie		Willie	Regan
Carley	Claude	Cammy	Maddy	Justin	Jo		Willy	Reggie
Cat	Cody	Carlin	Marietta	Karel	Jodi			Rene
Chrissy	Constantine	Carmine	Marion	Kyle	Jodie			Rey
Christie	Corey	Carroll	Meredith	Leland	Jody			Ricki
Christy	Corrie	Cary	Meryl	Marlo	Jude			Rickie
Clair	Cory	Caryl	Millicent	Martie	Kellen			Rikki
Claire	Dale	Casey	Nichole	Marty	Kelly			Robin
Clare	Daniel	Cass	Nickie	Michal	Kim			Ronnie
Connie	Dannie	Cecil	Nicky	Mickie	Kip			Sal
Courtney	Danny	Clem	Niki	Micky	Kirby			Sam
Dory	Darryl	Clemmie	Nikki	Perry	Kit			Sammy
Emmy	Daryl	Cris	Page	Phil	Kris			Sandy
Erin	Demetris	Daffy	Paige	Pooh	Lane			Sascha
Esme	Denny	Dallas	Patrice	Quentin	Lanny			Sasha
Evelyn	Devin	Dana	Patsy	Randy	Lee			Scotty
Felice	Devon	Dani	Pattie	Ray	Leigh			Shay
Georgia	Dorian	Darby	Patty	Ricky	Lesley			Shayne
Ginger	Drew	Darcy	Pennie	Robbie	Leslie			Shea
Grace	Eddie	Deane	Penny	Ronny	Lin			Shell
Gretchen	Eddy	Del	Randi	Sean	Lind			Simone
Haleigh	Edie	Dell	Randie	Shane	Lonnie			Sonnie
Haley	Francis	Dennie	Rory	Shaun	Loren			Torey
Hilary	Frank	Dion	Ruby	Shawn	Lorne			Val
Hillary	Fred	Dionis	Saundra	Sonny	Lou			Vale
Holly	Freddie	Dominique	Sayre	Tate	Mattie			Valentine
Ira	Freddy	Donnie	Shaine	Ted	Matty			Virgie
Isa	Gay	Donny	Shannon	Teddie	Maurise			Winnie
Isador	Gayle	Elisha	Shelby	Teddy	Max			Winny
Isadore	Gene	Fran	Shelley	Terry	Maxie			Wynn