Linguistic data of 32k film subtitles

Harshal Shinde, Amogh Inamdar, Tanishka Adhlakha, Yash Gaikwad

02/12/2023

Contents

1. Problem Statement

Televisions have been a major source of our entertainment since a long time. We watch movies regardless of the language or countries they are released in. They subconsciously make us live the feelings the character passes through in a movie. Today there has been a large surge in K-Drama viewership around the world even when they don't understand the language, subtitles play a vital role in such times. Subtitles also help when there are movie characters who use heavy accent.

Robertjoellewis-film-subtitles gives us a list of extensive data related to movies having information such as movie names, when they were released, their runtime, ratings, count of words the movie had along with sub categories of positive, negative words also count of nouns and pronouns used. The data set can be used to find how different types of words used in various genre create a better impact and can get better rating. Reseaching on such data will provide a good insight for further movie producers on when and words should be used. Use of positive words create a sense of light hearted movie which attracts large audience and get above average ratings where as serious movies like documentory and biography are where people download subtitles to listen to words more carefully, also in action movies people use subtitles as there is a lot happening and it is hard to listen everything.

2. About the Dataset

The dataset majorly divides the movies on the basis of genre. It gives extensive information on various factors like: -Movie Year -Total Words -Harm Virtue -Rating rank -Count of positive words used -Count of negative words used -Subtitles download count

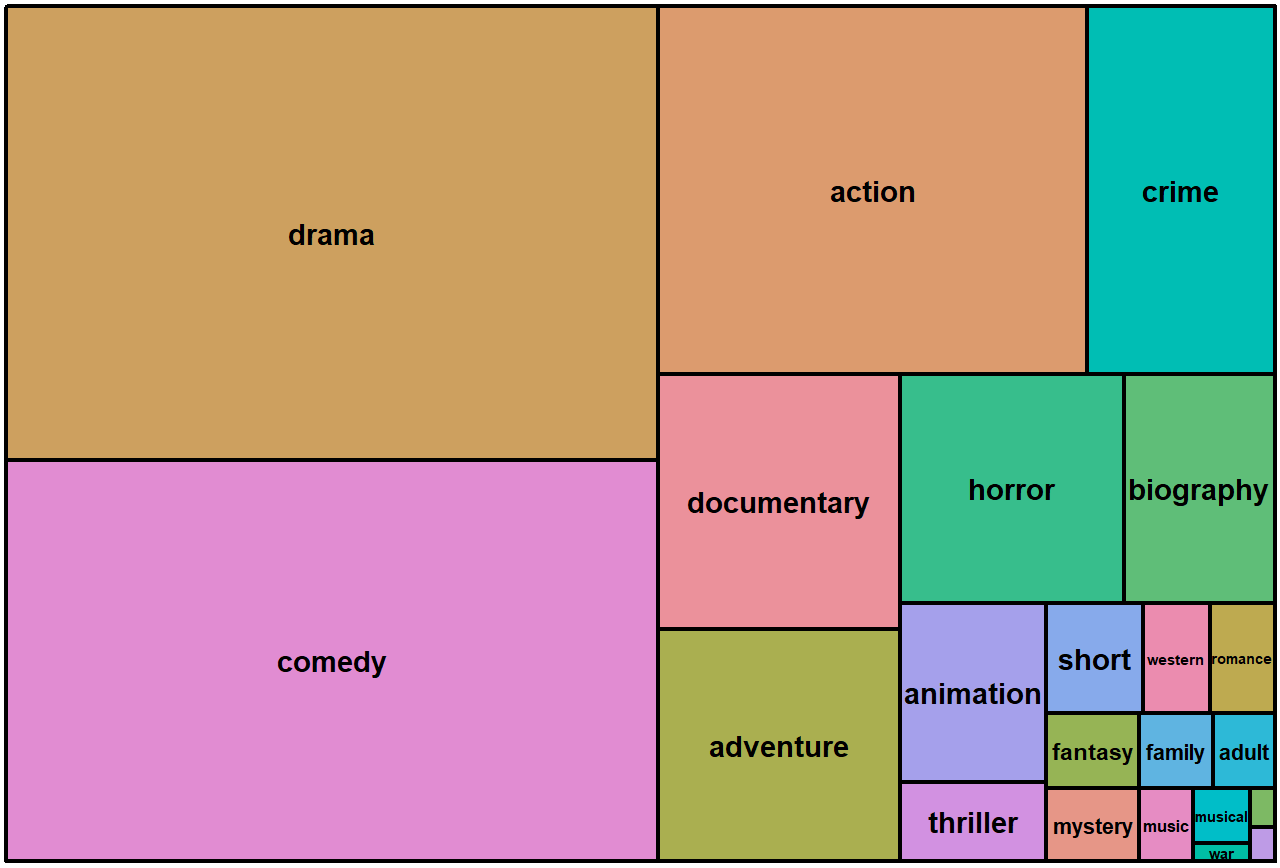
The dataset is referred from "https://data.world/robertjoellewis/film-subtitles (https://data.world/robertjoellewis/film-subtitles)".

3. Output

3.1 Distribution of movies

How are all the movies distributed along the years on the basis of genre?

Fig 1: Treemap of Category Frequencies

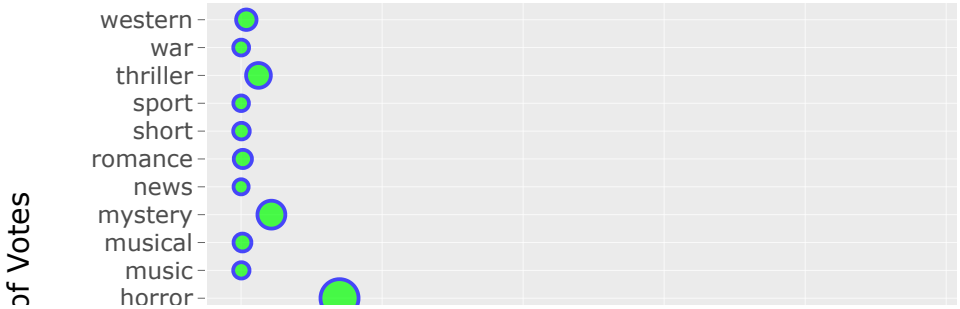


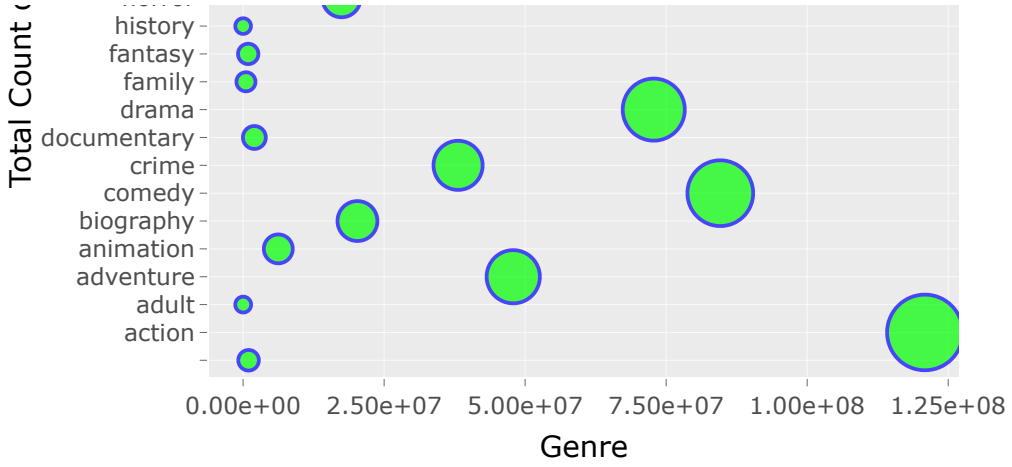
From the treemap plot we can see that drama and comedy genre movies equally dominate, followed by other genres like action, adventure, horror, crime and documentary. Other genres movies occupy a small space in comparison to drama and comedy movies. We can understand that audience would show much interest in these 2 genres resulting in them being much more produced.

3.2 How does rating works with votes?

How many people did vote for the movie based on genres?

Fig 2: Genre vs Total Count of Votes

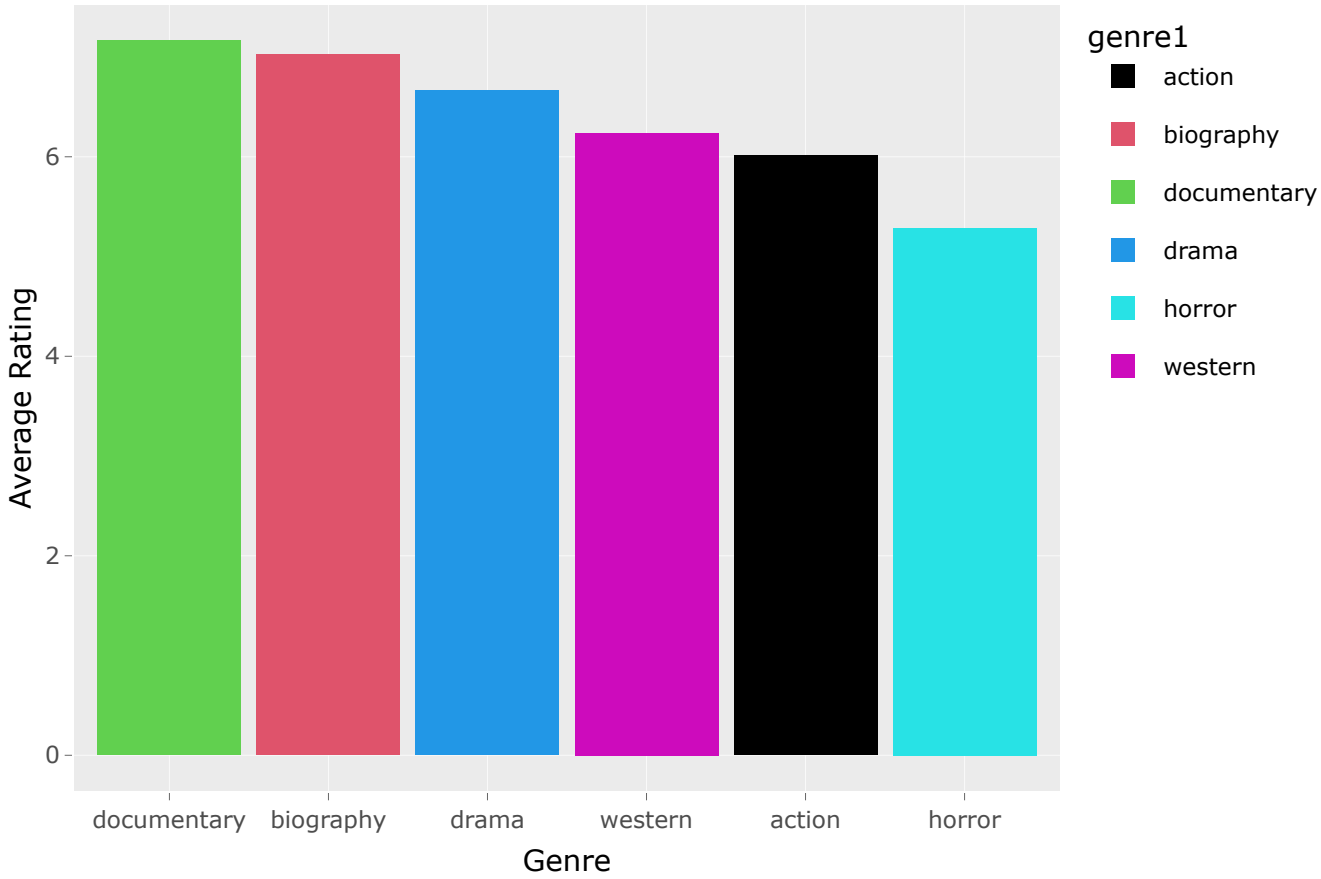




We can observe from the bubble plot that there were many votes given to action, comedy and drama movies, whereas other people did not vote for other movies. Action movies excites people and attracts more audience as much as comedy movies which are light hearted and fun to watch. A key detail that we can see from bubble plot and previously drawn tree map is that even though the action movies got more votes even though they are made fewer than drama and comedy genre movies. This gives us an important insight that people prefer comedy movies more.

What are the average ratings of each genre? Does more votes equal to higher rating?

Fig 3: Genre vs Average Rating



The plot shows a bar graph where x-axis represent various movie genres and y-axis represent average_rating. We come to know a whole new fact that one of the fewer produced and voted genres like documentary and biography have the highest average rating crossing 7 on the other hand popularly produced and rated genres like

drama have low average rating. Audience thus like to watch documentary and biography genre movies but being less produced gets less votes but on an average all movies have good ratings, whereas in case of drama genre movies have a variety of ratings including less rating resulting in lower average rating.

What are the average ratings of genres along the years?

Fig 4: TimeSeries Analysis for Documentary VS Drama

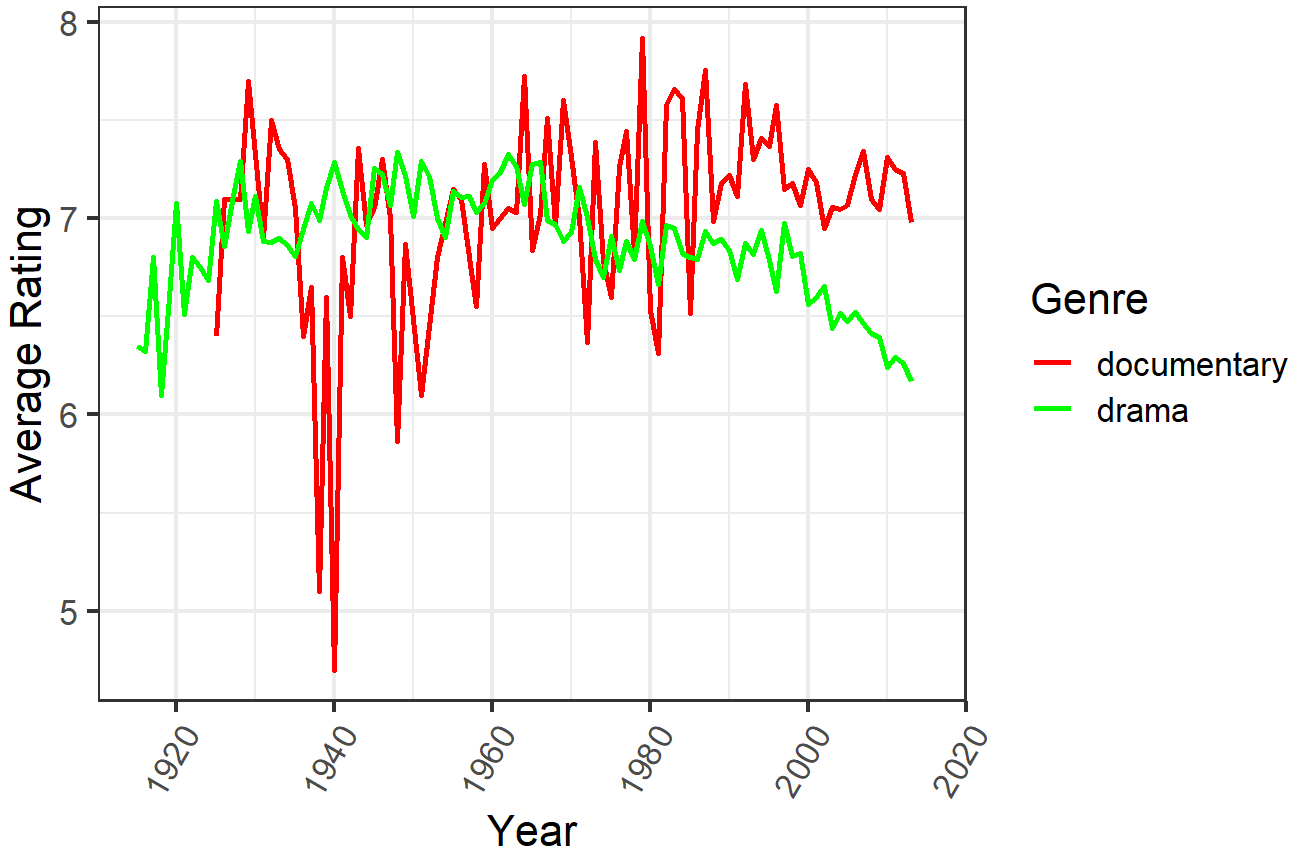
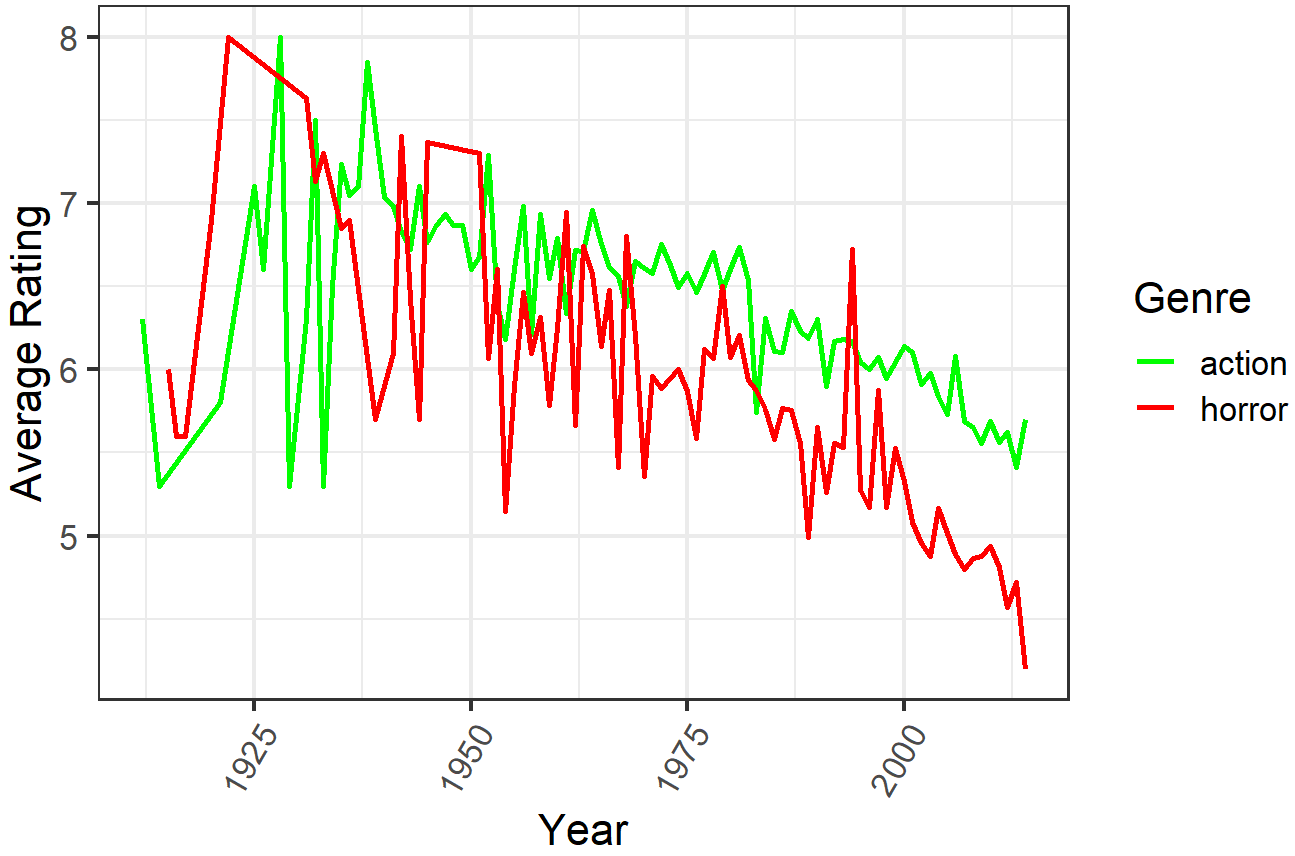


Fig 5: TimeSeries Analysis for Action VS Horror

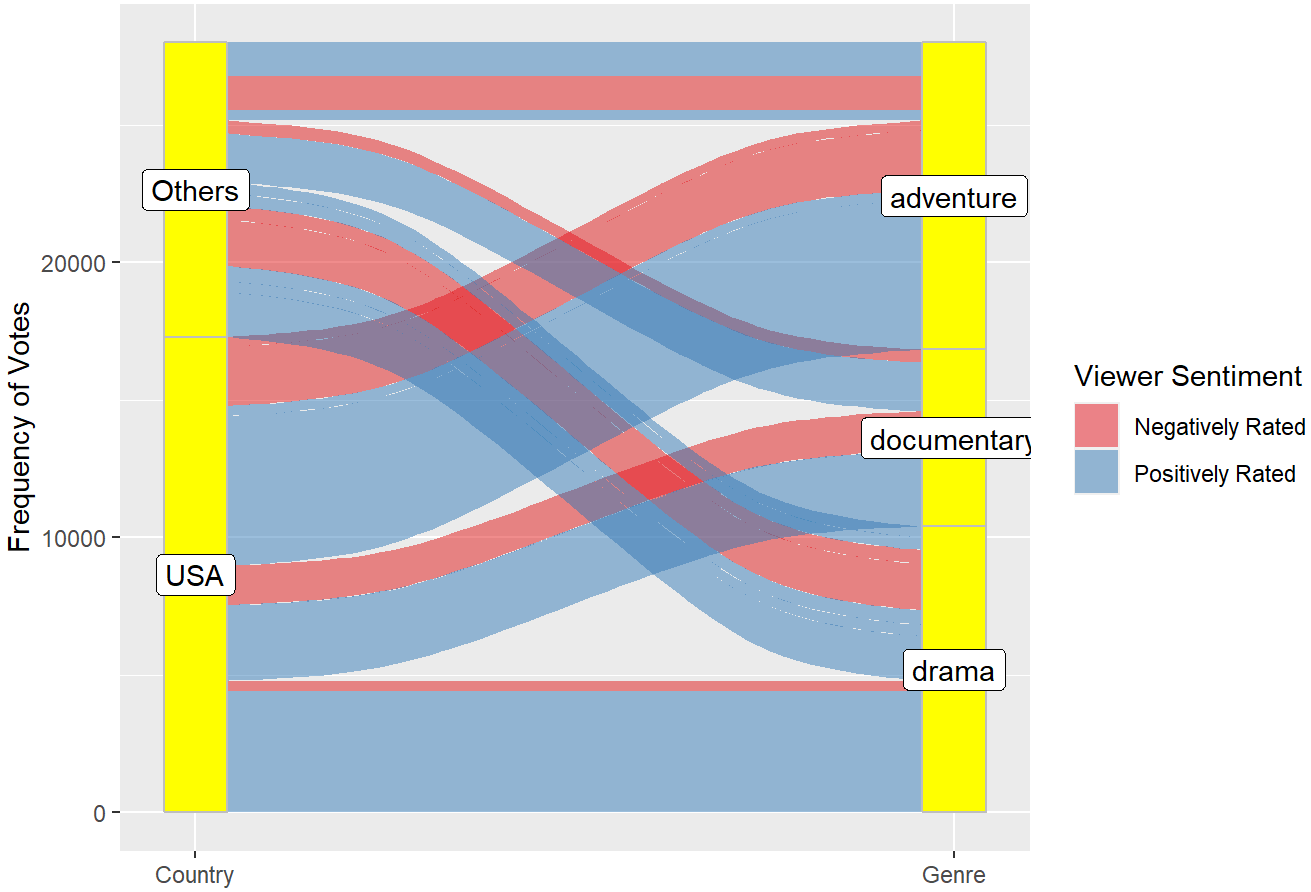


Diving deep into the ratings in specific genres we can see from the above plots that movies from action and drama genre have a decline in rating over the past years and dropping under 6, this suggests us why the average rating for these genres have declined. On the other hand the rating for Documentary movies have stayed pretty much steady around the range of 7 to 7.4 keeping the average rating high. Also there are some genre's with a high variable ratings over time such as horror movies as shown in line graph above.

3.3 How does rating works with emotions?

Are genre's rated positively and negatively equivalently in USA and rest of world?

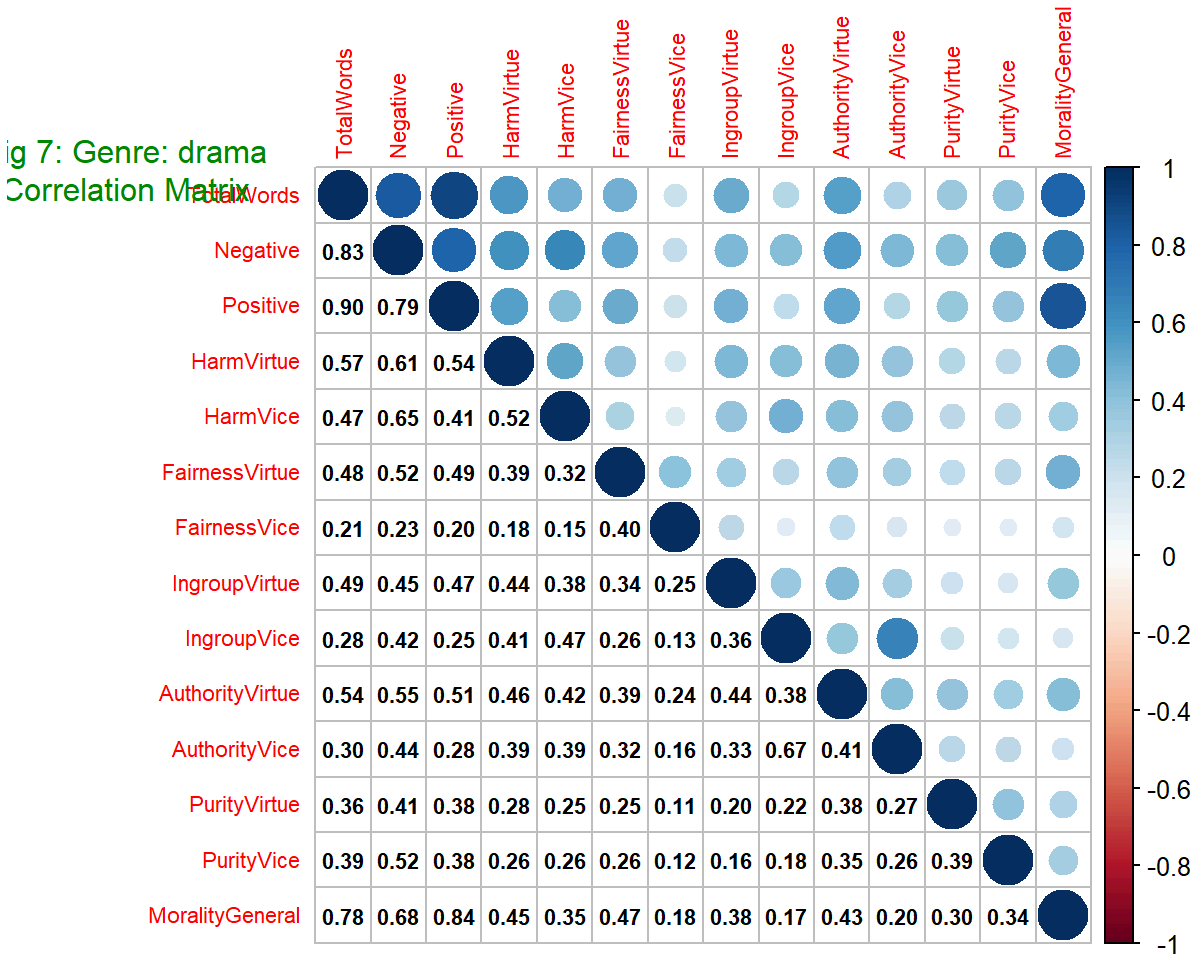
Fig 6: Distribution of Sentiment over Countries



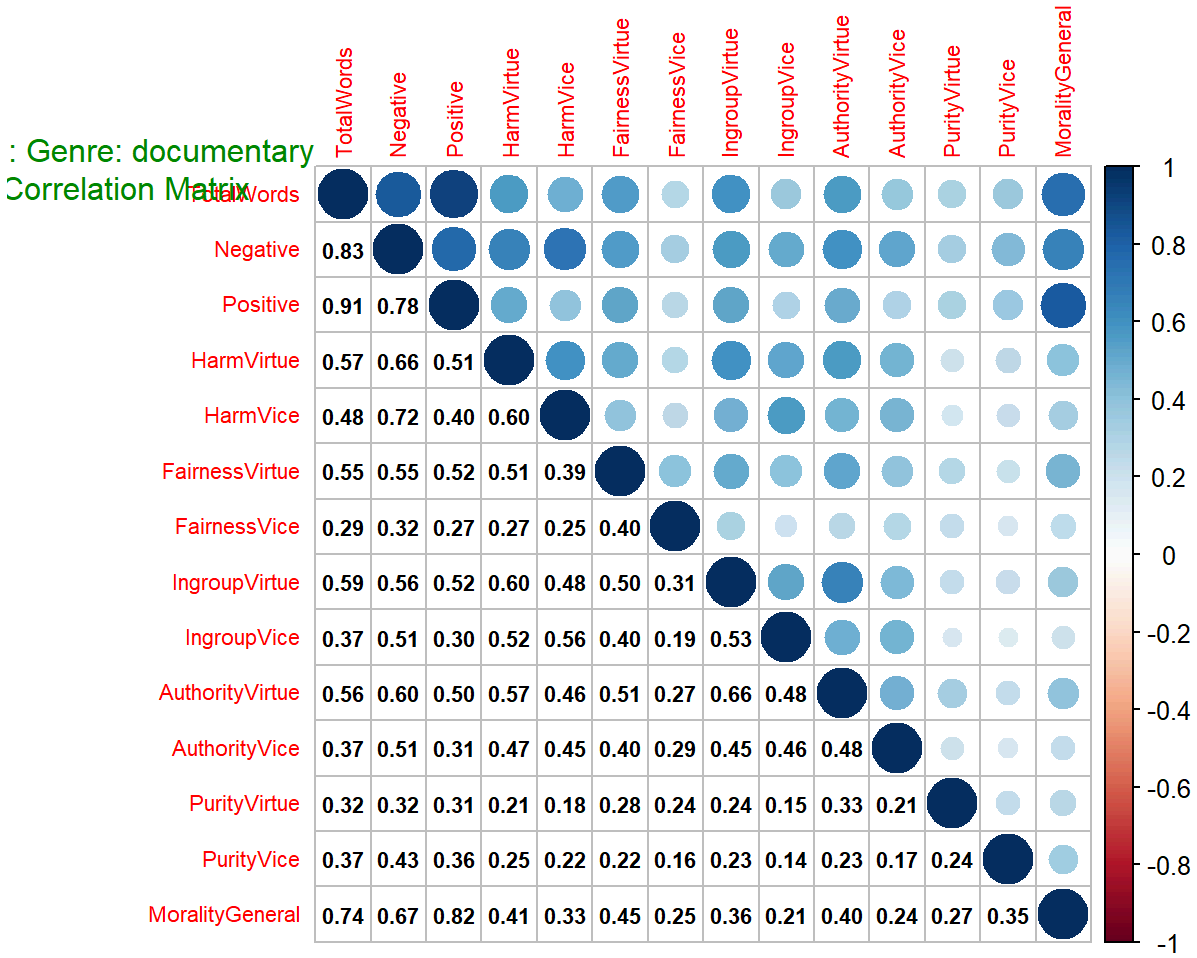
The above alluvial plot shows country and genres assigned to vertical axis that are parallel. Vertical sizes of the blocks are proportional to the frequency, and so are the widths of the alluvial. The relation between country and genre based on negative rating and positive rating can be seen. USA has more number of positive ratings throughout all genres. Whereas other countries have equal positive and negative ratings for adventure genre and more positively rated documentary movies and more negatively rated drama movies.

How emotions and words are related?

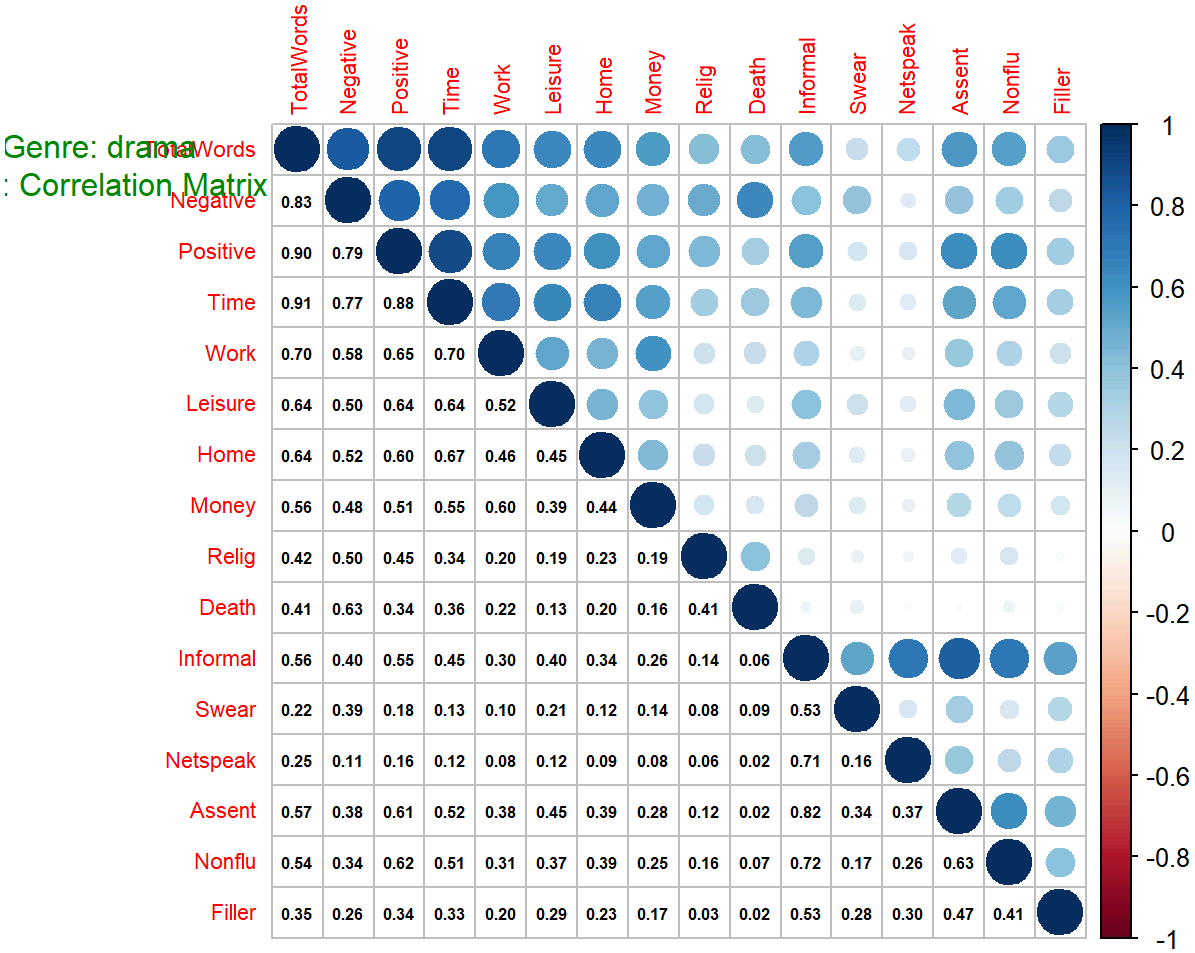
ig 7: Genre: drama
Correlation Matrix



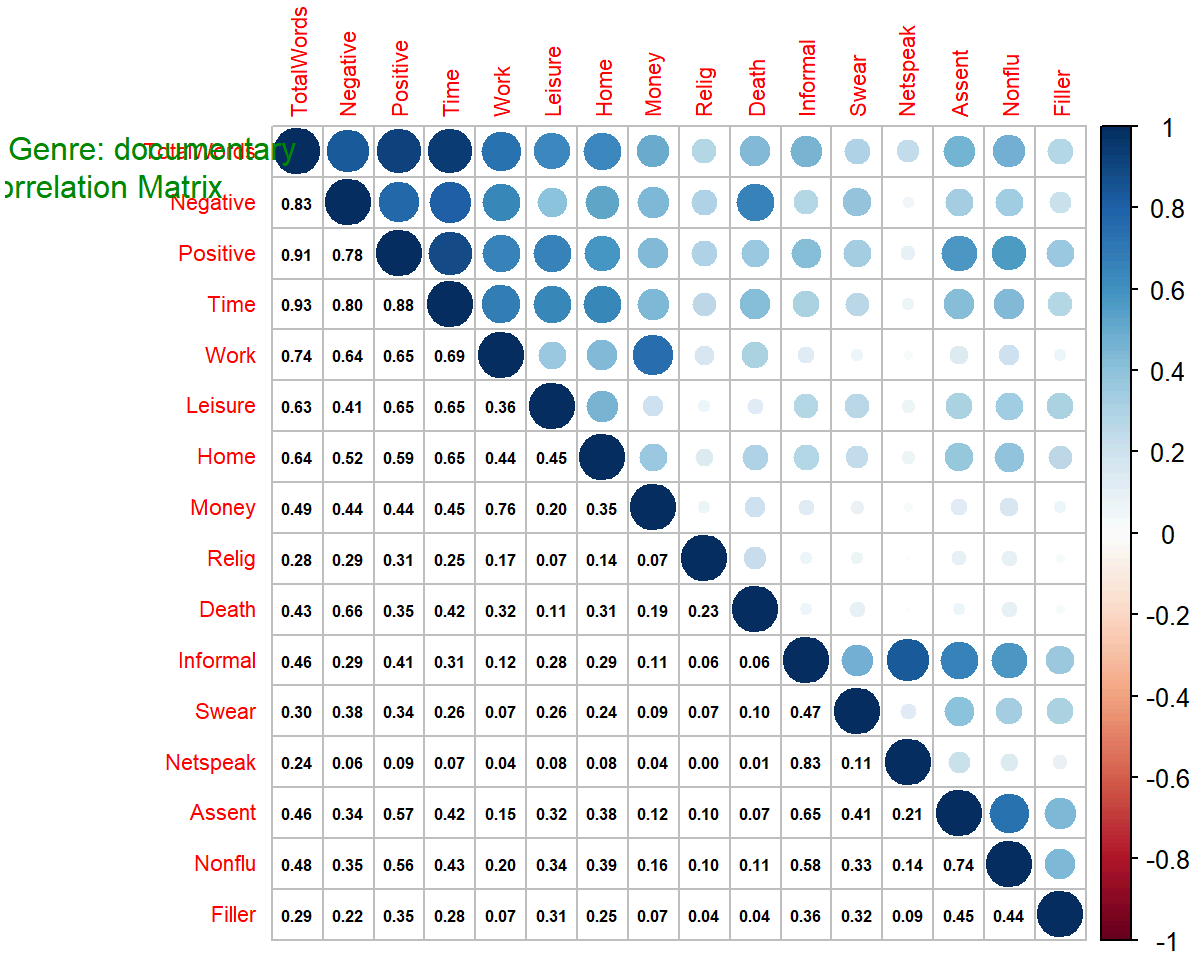
: Genre: documentary
Correlation Matrix



Genre: drama
: Correlation Matrix



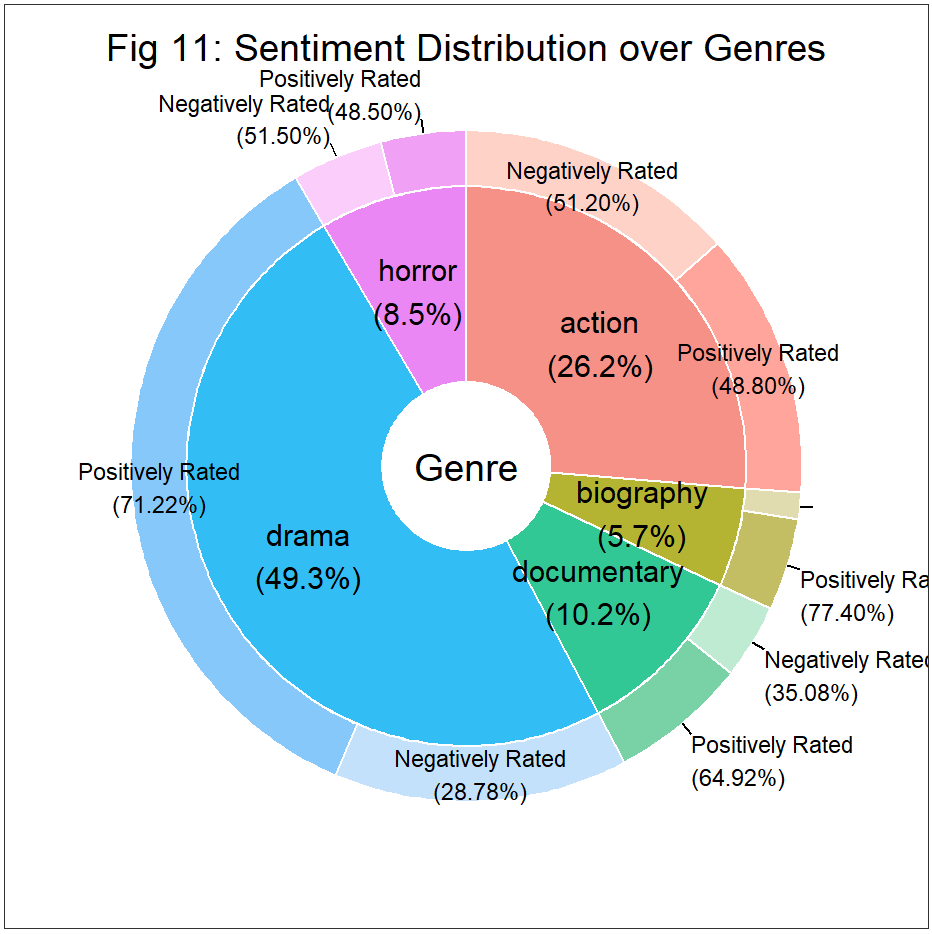
Genre: documentary
orrelation Matrix



The above correlation plot is used to highlight the most correlated words and emotions in a data table. We can see each variable fits always perfectly correlates with itself. The matrix is symmetrical along the diagonal where lower diagonal shows the numerical correlation and upper represents the depth in color and shape representation. Time and Positive words have high correlation between them in Documentary genre of 0.88 which represents both the words have a great direct relationship whereas a poor relationship can be seen between Filler and money of 0.04 which shows that they have least dependency between them. Also we can see a 0 value in Netspeak and Relig which denotes there is no relation whatsoever in them.

Similarly the correlation between emotions can be seen in above graph for Documentary and Drama genre.

How sentiments and words are related?



The distribution of positive and negative sentiments can be seen over various genres. The count of positive sentiments are compared with negative sentiments to get how much area is covered by them respectively in each genre. We can see the Drama genre consists highly on positive sentiments creating happiness subconsciously for the viewers, on the other hand we can see horror and action movies have equal distribution of positive and negative emotions. This creates an emotional roller coaster for the viewers.

4. Conclusion

Over the analysis we learned how movies interact with viewers. The study shows graphical representation on how genres shows variations with various scenarios,

1. The distribution of various genres of movies was plotted to find that there is a large number of Drama, Comedy and Drama movies produced in comparison to all other genres, Drama and Comedy being equal to all the remaining genres.

2. The total votes viewers gave to the movies on the basis of genre revealed that Drama movies even though being the most produced was not the one with most number of votes, whereas Action and Comedy movies received more votes.

3. When the average ratings for genre was calculated we observed that Documentary movies had the highest rating, this revealed that more number of votes doesn't necessarily means higher the average rating.

4. The reason for higher average rating for Documentary movies was found when the time series line graph was drawn for Documentary and Drama genre showing the decline for ratings of Drama movies over past few years.

5. In depth analysis of frequency of words with viewers sentiments in USA and other countries showed Positive and Negative rating given by the viewers. This showed that viewers from USA have given more positive ratings to all genres whereas other countries have mixed ratings according to genres.

6. Further how words and emotions are related to create a sense of belonging for viewers was seen using the corrplot and which words create a positive impact was found.

7. Viewers sentiments were seen using the distribution of positive and negative sentiments over genres which showed Drama movies have high positive sentiments compared to other genres.

5. References

1. Robert Joel Lewis, Matthew Grizzard, Sydne Lea, Doug Ilijev, Jin-A Choi, Lisa Müsse & Gabriela O'Connor (2017) Large-Scale Patterns of Entertainment Gratifications in Linguistic Content of U.S. Films, Communication Studies, 68:4, 422-438, DOI: 10.1080/10510974.2017.1340903

2. IMDB Movie Ranking. doi:https://www.imdb.com/search/title/?groups=top_250&sort=user_rating (doi:https://www.imdb.com/search/title/?groups=top_250&sort=user_rating)

3. Seih, Y. T., Chung, C. K., & Pennebaker, J. W. (2011). Experimental manipulations of perspective taking and perspective switching in expressive writing. Cognition & Emotion, 25, 926–938. doi:10.1080/02699931.2010.512123 (doi:10.1080/02699931.2010.512123)

4. Tamborini, R. (2011). Moral intuition and media entertainment. Journal of Media Psychology, 23, 39–45. doi:10.1027/1864-1105/a000031 (doi:10.1027/1864-1105/a000031)

5. Weber, R., Popova, L., & Mangus, J. M. (2013). Universal morality, mediated narratives, and neural synchrony. In R. Tamborini (Ed.), Media and the moral mind (pp. 26–42). New York, NY: Routledge.