Non-English Film Engagement on IMDb

Emily Carpenter Marcus Almanza Janet Matthews-Derrico
Yoshie Bell-Souder

1 ABSTRACT

As media is increasingly globalized, the success in recent years of films like *Parasite* (Bong, 2019) and *All Quiet on the Western Front* (Berger, 2022) demonstrate the widespread potential for non-English language films in the English-speaking market. First launched in 1990, today the Internet Movie Database (IMDb) is a comprehensive, U.S.-based online database of information primarily pertaining to films and television shows. Its reviews and ratings systems offer a way to measure user engagement, which can represent a mode of success that is different from more traditional measures such as ticket sales and awards.

With this project, we aimed to explore user engagement with non-English language films listed on IMDb. We were most interested in looking for interesting patterns in the number of user votes—that is, the number of distinct user profiles that have voted on a film rating between 1 and 10—as well as the film ratings themselves. To that end, we developed a set of specific questions around the number of votes, ratings, film language, release year, and genre that we hoped could be answered by the data.

We began by preprocessing and integrating several different datasets available from IMDb. From there, we developed tools to fill in missing data primarily centered around the film language—an integral part of our research questions. We then employed data mining tools in an effort to determine whether there are certain conditions that lead to greater user engagement on IMDb. We found that while there is overall a heavy disparity between English and non-English films, the language of the film itself as well as the genre may have a sizable effect on user engagement. These findings could shed light on which non-English films should receive a greater investment when it comes to U.S. distribution as well as how they should be listed on IMDb.

CCS CONCEPTS
• Applied computing → Arts and humanities • Information systems → Data management systems • Information systems → Data mining

KEYWORDS
Data mining; Database systems; International films; movies; IMDb

ACM REFERENCE FORMAT:
Marcus Almanza, Yoshie Bell-Souder, Emily Carpenter and Janet Matthews-Derrico. 2023. Non-English Film Engagement on IMDb: *ACM, Boulder, CO, USA, 13 pages.*

2 INTRODUCTION

There is a lack of research on the relationship between film languages and their success, whether that be measured by star ratings, box office income, or the number of ratings. Our hypothesis was that non-English language films likely provide a plethora of high-quality movies that are less likely to be watched because of their original language. In 2020 when Bong Joon Ho accepted an Academy Award for Best Foreign Film for *Parasite*, he said “Once you overcome the one-inch tall barrier of subtitles, you will be introduced to so many more amazing films.”

We developed six key research questions for this project:

- Do non-English films have fewer reviews and ratings, regardless of their average rating?
- Do international films have higher ratings due to fewer view counts?
- Do certain languages do better when it comes to ratings/reviews overall?
- Do certain genres of non-English films do better? How does this compare to English film genres?
- Which films are "underrated gems," those that have high ratings, but relatively low view counts?
- Do certain directors have more user-engagement than others, and does the rate of engagement vary based on the language?

Our research revealed patterns within non-English film engagement that shed light on the willingness of users to watch content from countries outside of their own. These questions in particular showcase the differences between English and non-English films and the ways users of IMDb interact with them.

Since the Second World War, English films have become a dominant force in international markets and continue to have an increasingly large share of the international film market. By identifying the factors that contribute to the success of films in international markets, this study can inform filmmakers, distributors, and other industry stakeholders about the factors that influence the popularity of films, as well as help viewers find their next favorite movie.

3 RELATED WORK

The IMDb dataset has been used a number of times, primarily for exploratory analyses conducted on older iterations of the dataset. An exploratory data analysis on IMDb movie titles searching for overall trends in popularity was performed and can be accessed here: https://www.kaggle.com/code/slayomer/eda-on-imbb-film-dataset. Another exploratory analysis focused specifically on the popularity of Netflix titles on IMDb, can be accessed here: https://www.kaggle.com/code/keswanirohit/netflix-visualization-and-eda. Finally, the IMDb dataset was also used to create a movie recommendation system for users: https://www.kaggle.com/code/jasonlei0420/ds5230-movie-recommendation-system. Our presented research here was conducted on a newer iteration of the dataset, and focused on information not previously explored (i.e. English vs. non-English films).

4 DATA SET

Our data sets can be downloaded from this URL: https://datasets.imdbws.com/, and the documentation can be found here: https://www.imdb.com/interfaces/. We used the following data sets from IMDb: title.basics.tsv.gz, title.ratings.tsv.gz, title.akas.tsv.gz, title.crew.tsv.gz, title.principals.tsv.gz, and name.basics.tsv.gz. The other available data set relates to television episodes, which is outside the scope of this research project.

IMDb houses data for over 10.1 million titles and almost 630,000 films. Our first data set, the ratings data set, contains just 3 data points: "tConst" (the unique identifier of a title), an average rating, and the number of unique votes.

Our second data set, title.basics, has 9 attributes, including tConst. Using this data set, we were able to start the data reduction process using the following attributes: isAdult (a boolean to mark if the film contains adult content) which we used to remove all explicitly adult content, titleType (such as TV show, episode, film, short film, etc.) which we used to eliminate non-feature films from the data set. Here we also have the bonus of having an endYear attribute, which in theory should only be used for television shows and be empty for everything else. This helped us catch errors. When researching patterns, we used this data set to access genres of films, the year the film was released, and the runtime.

The third data set we accessed is the title.akas (also known as) data set. This includes the language of the film and the regions in which the original title is used. It does not, however, tell us which language is the original language. For example, a film dubbed in multiple languages will have multiple entries in title.akas–one for each of those dubbed languages. To determine the original, we used an API and an open-source Python library, which will be elaborated on in the following section.

Upon further research, the datasets title.crew, title.principals, and name.basics were also added to this project. A combination of these datasets was used to get the name of the director(s) for each film. This was not originally planned research, however, once we were able to learn more about the diversity within different languages, we saw the need to explore this further. For these three datasets, the only attributes used were the "nConst" - the unique identifier for individuals, the name of the director, and the tConst. This was then simply added to our larger data warehouse.

There were limitations posed by this dataset. For instance, title.crew only contained the tConst and the nConst of the director, but not the director's name. We also discovered that there is no way to connect the original language to the English title without using outside tools. The datasets available on the IMDb website for public use also lack information that we were originally hoping to obtain, such as U.S. release dates, country of origin, worldwide box office gross, and budget.

5 MAIN TECHNIQUES APPLIED

5.1 Data Preprocessing

Since we were using multiple files from the dataset we had found, our first step for data preprocessing was data integration. For this step, we determined which columns from each file were relevant to the questions we were trying to answer (as previously stated). After determining the needed columns, we used the common movie id amongst all of the files to combine them into a single dataset to reduce any unneeded jumping between files and to more easily process our data.

After integrating the files into one the next step of our preprocessing was data reduction. The first part for this was to select the data points that corresponded to movies. To accomplish this, we utilized the title type column to remove all of the non-movie data objects (this included shorts, tv series, etc). The second part of our reduction was to use the isAdult column to remove films that are considered to be explicit adult movies. All values equal to one in that column were removed.

The final step for preprocessing that we did was data cleaning. While going through the initial dataset we noticed that the movie language column contained all of the available languages for the corresponding movie, and not just the initial language in which the movie was released. To fix this issue we utilized two Python libraries (TheMovieDatabase aka TMDB and Cinemagoer) to go through our integrated file and determine each movie's initial language by using the api from those libraries to get the correct and relevant information.

5.2 Data Warehouse

Using all the techniques from the preprocessing to go from multiple files into a single condensed file was how we created our data warehouse. With our data warehouse we were able to go through all the data that we needed instead of having to navigate through unnecessary points.

Since our data was a subset of a larger one with multiple files we modeled the warehouse as an independent data mart.

5.3 Data Modeling

Our research questions determined how we explored our dataset, and how we related each attribute. First, we set out to find a correlation between continuous variables using a correlation matrix chart with the Python scikit-learn library. The relationship between the Number of Votes and the Average Rating was 7.4%. As demonstrated by the correlation matrix below, there is only a bit of a correlation compared with other correlations. This result was able to focus more on one of our questions, do non-English films have fewer reviews and ratings, regardless of their average rating?

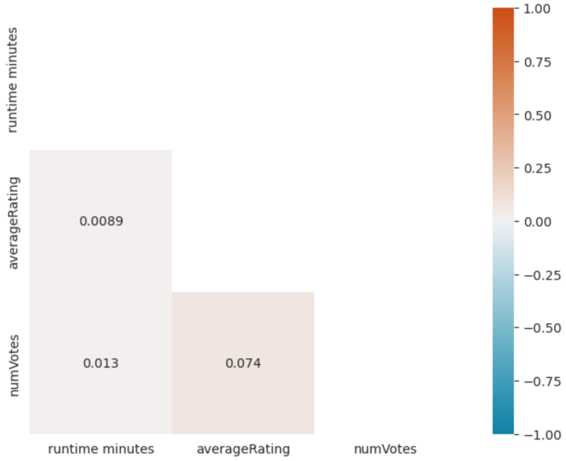


Figure 1: Correlation Matrix

We also could use the results for one of our last questions: which films are “underrated gems,” that have high ratings, but relatively low view counts? First, we made a scatterplot between the Number of Votes and the Average Rating relationship using the Python matplotlib library. Then we made a hypothesis to make a slope from the dense data points area to the highest rating. To find underrated gems, we observed out-of-standard deviation ranges. It roughly found "underrated gems". However, when we looked at the scatterplot carefully around the rating of 8, the graph showed that there are some data points that have a huge number of votes–thus disqualifying it as an underrated gem. When the data points make a line, we may be able to use gamma regression instead of linear.

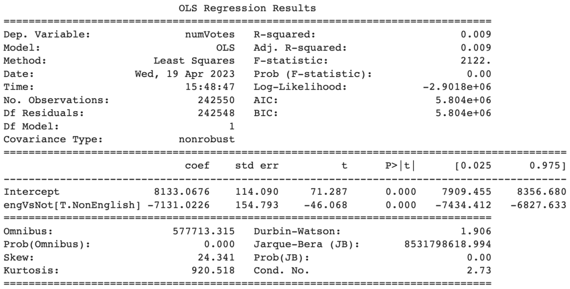
Next, we marked each data point between relationships in the number of films by language by release year, as well as the total number of films by release year using Tableau. We found that we could use linear regression for the number of films in each language by year, and exponential regression for the total number of films by year.

Lastly, we observed the genre of each movie with the Python scikit-learn library whether we can find K-means clustering or not. The smallest the K-means iteration, which was K = 3, was too widely distributed. In running further iterations, we were able to find different results, but they were too rough to be of use in our evaluations. This is probably due to the fact that many films are not assigned a single label, but instead have multiple labels, like Drama, Mystery rather than just Drama. It was hard to determine genre-based patterns using K-means clustering.

6 KEY RESULTS

6.1 Number of Ratings vs. Average Rating for English and non-English Films

Running a T-test, the non-English group of movies has significantly fewer votes (reviews). On average the non-English films had 7,131 fewer votes than English films (t value = 46.068, $p < 0.001$). This assumes that variance is roughly equivalent, we would say non-English films have fewer reviews, but the rating is not much affected by the fewer reviews.



```
                             OLS Regression Results
==============================================================================
Dep. Variable:               numVotes   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.009
Method:             Least Squares   F-statistic:                     2122.
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        15:48:47   Log-Likelihood:            -2.9018e+06
No. Observations:              242550   AIC:                         5.804e+06
Df Residuals:                  242548   BIC:                         5.804e+06
Df Model:                           1
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                8133.0676    114.090     71.287      0.000    7909.455    8356.680
engVsNot[T.NonEnglish] -7131.0226    154.793    -46.068      0.000   -7434.412   -6827.633
==============================================================================
Omnibus:                   577713.315   Durbin-Watson:                   1.906
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       8531798618.994
Skew:                          24.341   Prob(JB):                         0.00
Kurtosis:                     920.518   Cond. No.                         2.73
==============================================================================
```

Figure 1: T test for Number of Votes and English vs non-English films

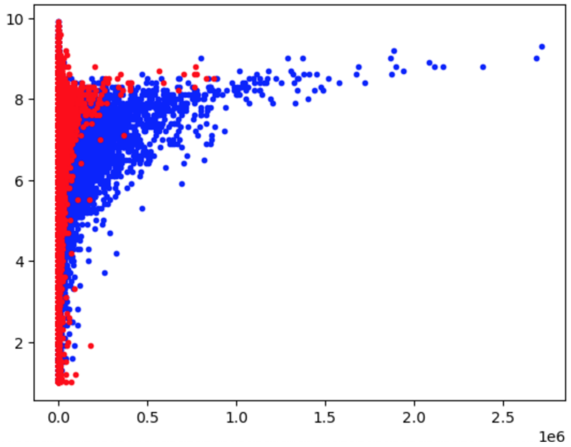
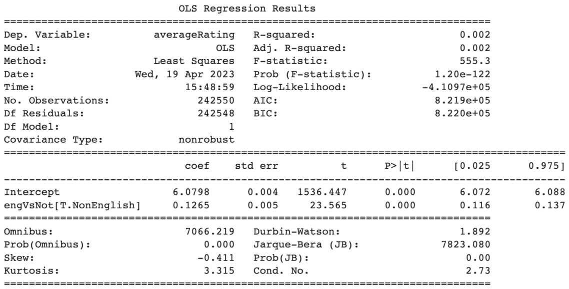


Figure 2: Scatterplot Between the Number of Votes and the Number of Rating
Blue: English films Red: non-English films

We then looked at the top five individual languages that had the most movies and had at least one hundred votes for the movie (Spanish, French, German, Italian, Japanese), and compared those to the average rating. The rating distribution was still consistent, and it seems like having more votes for a movie, overall, doesn't affect the rating too much after a certain point.

6.2 Individual Language Popularity

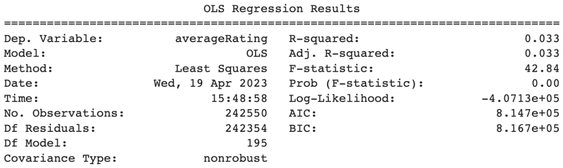
We again examined the p test, and found that non-English movie ratings on average (about 6.2) are higher than the average rating for English movies (6.0798). Looking at a standard error of 0.005, T value of 23.565, and P value less than 0.001. It is significant, but may not be practically significant.



```
                             OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:             Least Squares   F-statistic:                     555.3
Date:                Wed, 19 Apr 2023   Prob (F-statistic):          1.20e-122
Time:                        15:48:59   Log-Likelihood:            -4.1097e+05
No. Observations:              242550   AIC:                         8.219e+05
Df Residuals:                  242548   BIC:                         8.220e+05
Df Model:                           1
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   6.0798      0.004   1536.447      0.000       6.072       6.088
engVsNot[T.NonEnglish]      0.1265      0.005     23.565      0.000       0.116       0.137
==============================================================================
Omnibus:                     7066.219   Durbin-Watson:                   1.892
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7823.080
Skew:                          -0.411   Prob(JB):                         0.00
Kurtosis:                       3.315   Cond. No.                         2.73
==============================================================================
```

Figure 3: T test for Average Rating and English vs non-English films

Running an Analysis of Variation for average rating by languages, there are significant differences between the averages. (F = 42.84, p<0.01)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.033
Model:                            OLS   Adj. R-squared:                  0.033
Method:                 Least Squares   F-statistic:                     42.84
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        15:48:58   Log-Likelihood:            -4.0713e+05
No. Observations:              242550   AIC:                         8.147e+05
Df Residuals:                  242354   BIC:                         8.167e+05
Df Model:                         195
Covariance Type:            nonrobust
```

Figure 4: ANOVA test for Average Rating and Languages

In looking at the number of films for each language (excluding English), the top five were Spanish, French, German, Italian, and Japanese. Of these, only Spanish and French are among the top 10 most spoken languages in the world as of 2022 according to Ethnologue. The top five languages ranked by the number of user votes were French, Hindi, Italian, Spanish, and Japanese. This aligns slightly more with the most spoken languages in the world in that Hindi is number three after English and Mandarin Chinese. However, our results overall suggest that the number of user votes is not strongly tied to the number of global speakers of the film's language.

We then looked at individual languages that had at least 1,000 movies to see if any language seemed to have a significantly higher average rating. We found that Bengali had an average rating of about 6.89 among films that contained at least 100 votes. The standard deviation among average rating is 1.267. This information taken in conjunction with the overall dispersal of data suggests that the high rating among Bengali films is not simply a result of one or two films skewing the average. There are 272.7 million Bengali speakers around the world, making it the 7th most spoken language globally.

6.3 Genre

In examining the user engagement with regards to genre, we found that the most generally popular genre in non-English language films is Drama. We are defining popularity here by the total number of user votes rather than by rating–a high user engagement on a film with a lower rating can still be interpreted as a measure of success. The non-English popularity of the Drama genre is in contrast to English language films, in which the most popular genre was Action/Adventure/Sci-Fi. Although the genre composition for the top fifteen languages ('top' is again defined by the total number of user votes) varied from language to language, Drama was a notable percentage for all but one.

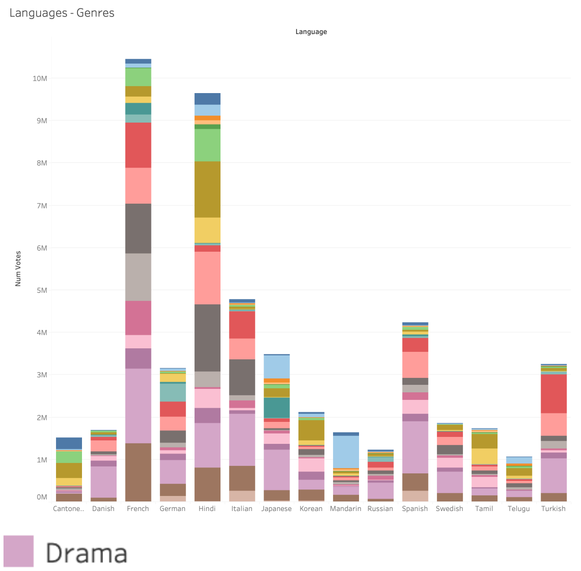


Figure 6: Number of films in each language by genre

Drama films comprised over 22% of non-English films in the dataset, and accounted for nearly 8.6% of all non-English film user votes. By comparison, the Drama genre accounted for only 11% of English film titles and 3.2% of overall user votes. Additionally, of the user votes on English language Drama titles, over 20% can be attributed to only three films: *The Shawshank Redemption* (Darabont, 1994), *Fight Club* (Fincher, 1999), and *American Beauty* (Mendes, 1999).

To further explore whether a non-English film that is classified as Drama has a higher chance of accruing more user votes, we set a limit at >500 votes and calculated the lift, which was 0.611. This value indicates that there is a positive correlation, but it is not as significant as the overall percentages mentioned above might have suggested. It is possible that non-English Drama films have a slightly higher chance of succeeding when it comes to user engagement over other genres because Drama represents a much smaller portion of English language films and there is less competition. It can also be argued there is still a perception that "foreign films" should be serious and somewhat staid, which aligns with the Drama genre.

While Drama accounted for the highest number of non-English films overall, we were also interested in discovering other measures of popularity that might equate to higher user engagement. With that goal, we identified Crime/Drama/Mystery as the genre with the highest average of user votes per film. This genre accounts for only 0.004% of non-English films and 0.47% of films overall. Of all Crime/Drama/Mystery films, over one quarter were released within the past ten years which may explain the popularity of a genre that makes up a relatively small percentage of films in the dataset. In contrast, the genre with the highest average votes per English language film was Action/Adventure/Sci-Fi.

It is important to note that the specificity of the Crime/Drama/Mystery genre also limits the number of films. If we expand it to include all films with “Crime,” in combination with “Drama,” “Mystery,” or “Thriller,” its overall percentage rises to 2.75%. However, part of the aim of this project is to explore what conditions may lead to higher user engagement, and the specificity of Crime/Drama/Mystery is one of those conditions.

As illustrated in Figure 7, the film language composition of the Crime/Drama/Mystery genre is relatively similar to the overall breakdown of non-English films by language.

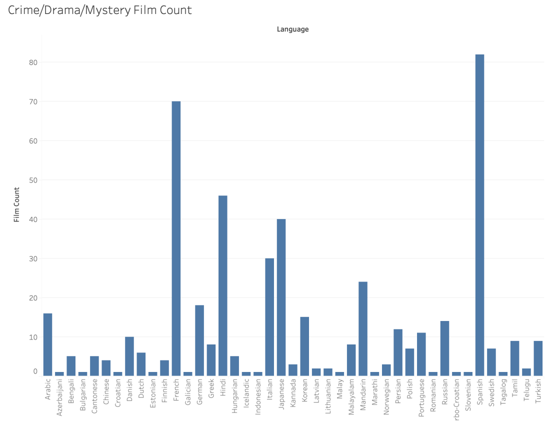


Figure 7: Crime/Drama/Mystery by number of films

Spanish and French have the most films, and the language with the third most Crime/Drama/Mystery films is Hindi. Interestingly, Hindi films in the genre had the most user votes overall, as well as the highest average of user votes per movie:

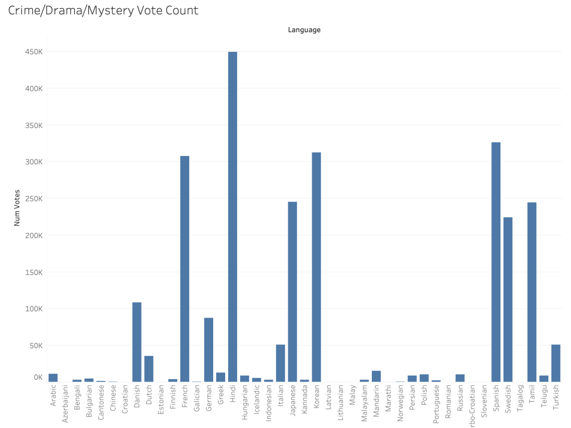


Figure 8: Crime/Drama/Mystery by user votes

The lift for non-English Crime/Drama/Mystery to accrue more than 500 votes is 2.25, which points towards a positive correlation that is substantially higher than that for the Drama genre. While this is perhaps suggestive of conditions for larger user engagement, using the Naïve Bayesian probabilistic classifier for Hindi Crime/Drama/Mystery films that have more than 500 user votes (the highest average/film) we are left with only 5.313 x 10^{-5}. This calculation is reflective of the size of the subset relative to the large number of data points in the full dataset, and indicates that success within this genre should be taken very cautiously.

We found that within the Crime/Drama/Mystery subset of the overall dataset it was important to examine the language breakdown as much as possible to account for outliers. For example, Swedish language films had the seventh highest number of user votes, but over 97% of them were for a single movie—*The Girl with the Dragon Tattoo* (Oplev, 2009). This outlier indicates another important condition for success: previous English language crossovers. *The Girl with the Dragon Tattoo* is based on a Swedish novel by the same name by Stieg Larsson and part of a trilogy that had sold over 17 million copies in the U.S. by 2011.

Overall, our key findings in relation to genre indicate that Drama accounts for both the highest number of user votes and highest number of films in our non-English dataset, especially as compared to the English language film dataset. While it does not guarantee user engagement, it is an overall positive condition. Crime/Drama/Mystery is a much smaller portion of the dataset. However, its increased output in recent years combined with the above calculations regarding user votes point to it as a genre that is worth keeping an eye on in the future.

Finally, looking at genres over the years it seems like for both English and non-English there was a shift in overall popularity from Drama to Documentaries around the 1990s for English movies and 2000s for non-English movies. In this case popularity is defined as a certain genre containing at least 50 movies and each movie having at least 25 votes towards their ratings. This change is likely due to digital films, which made it easier to produce documentaries in particular in addition to giving filmmakers the ability to experiment more due to the cheaper cost. It can also be argued that documentary movie styles/layout changed to reflect more mass-market appeal around this time as well.

6.4 Underrated Gems

We sought to identify "underrated gems," films with relatively few vote counts despite a comparatively high average ratings. As mentioned in section five, Main Techniques Applied, we explored the relationship between number of votes and the average rating of films using a scatter plot, which slopes from the densest data point area on the left to the highest ratings on the right. Using a regression model with 2.5 upper and lower standard deviation to focus on the underrated gems, we were able to find about 400 films as underrated gems with about 1% confidence. A few of the results from films that average 2.5 standard deviations above the average when controlling for the number of votes are:

- *The Silence of Swastika* (Bhardwaj, 2021), a Hindi movie with 10,353 votes and a 9.5 rating
- *Taxi* (Sajja, 2023), a Telugu movie with 1,504 votes and a 9.8 rating
- *Divorce* (I G, 2023) a Malayalam film with 601 votes and a 9.6 rating
- *Svet Koji Nestaje* (Lalovic, 1987), a Serbo-Croatian documentary with 349 votes and a 9.5 rating.

56 films of the top 100 were Indian-language films. Part of this can be attributed to the fact that many of the languages of the Indian subcontinent are amongst the most spoken in the world, but also because English is a national language of India, making it widely spoken there. This increases the likelihood of Indians using IMDb, as opposed to other countries where English fluency is not the standard.

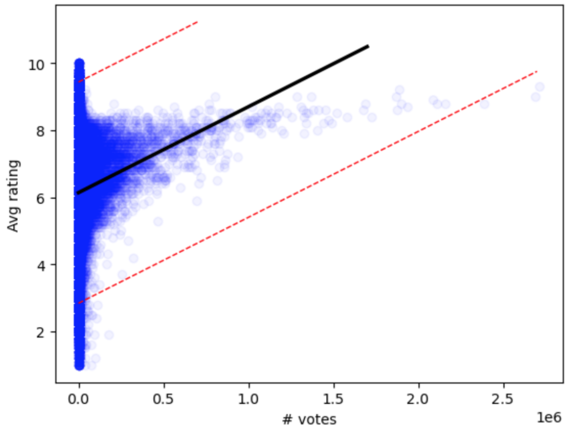


Figure 9: Linear regression Between the Number of Votes and the Average Rating with 2.5 standard deviation.

It can be hard to conclude with certainty that recent movies or movies with a relatively low number of view counts will continue to have such high ratings. If we expand this search and look at movies with an average rating of 8.0 or higher with more than 15,000 votes, but not more than 30,000 votes, we see another picture emerge when it comes to underrated gems. In the scatterplot below, we see about an even number of red and blue, red being films before 1979 and blue being films after 1979. This is surprising because the number of films made in the 2010s alone outnumbers the total number of films made before 1980. This would imply that the real underrated gems are actually older films.

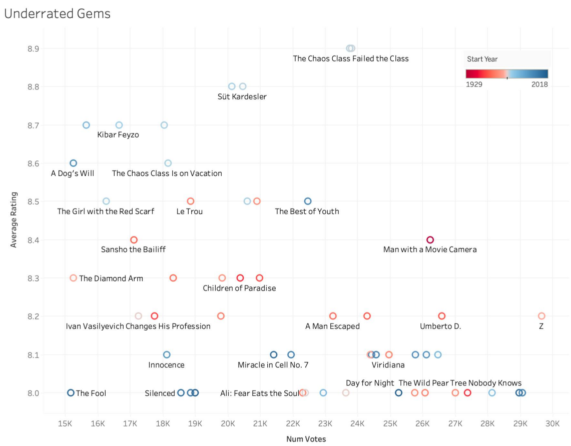


Figure 10: Underrated gems

Of the 74 films that fit the criteria above, only 21 of them are English-language, despite English films making up nearly half of the films on IMDb. This would indicate that a majority of underrated gems, those that are under-watched for their quality, are mostly older foreign films. A further measure of the quality of these films is the decennial poll produced by Sight & Sound, a magazine published by the British Film Institute, of the 250 Greatest Films of All Time. This list was last updated in 2022 and fifteen of the films we've determined are underrated gems also appear on this list. Nine of those are featured in both the Sight & Sound Top 100 and on our list:

- *Man with a Movie Camera* (Vertov, 1929), a Russian documentary ranked 9th with a rating of 8.4
- *Late Spring* (Ozu, 1949), a Japanese movie ranked 22nd with a rating of 8.2
- *Ali: Fear Eats the Soul* (Fassbinder, 1974), a German movie ranked 53rd with a rating of 8.0
- *Sansho the Bailiff* (Mizoguchi, 1954), a Japanese movie ranked 76th with a rating of 8.4
- *A Matter of Life and Death* (Powell & Pressburger, 1946), an English film with ranked 83rd with a rating of 8.0
- *The Leopard* (Visconti, 1963), an Italian film ranked 91st with a rating of 8.0
- *Ugetsu* (Mizoguchi, 1953), a Japanese film ranked 92nd with a rating of 8.2
- *Yi Yi* (Yang, 2000), a Mandarin film ranked 94th with a rating of 8.1
- *A Man Escaped* (Bresson, 1956), a French film ranked 95th with a rating of 8.2

6.5 Directors

Directors like James Cameron and Christopher Nolan are capable of amassing billions of dollars at the global box office throughout their careers. Steven Speilberg alone has managed to gross over $10 billion worldwide with hits like *Jaws* (1975), *Raiders of the Lost Ark* (1981), *Jurassic Park* (1993), *E.T.* (1982), and *Saving Private Ryan* (1998). Movie ratings for Spielberg's films alone account for 1.65% of all ratings for English films. In the same vein, Christopher Nolan's ratings are 1.74% of all English film ratings. Only five other English-language directors cross this 1% threshold: Quentin Tarantino, Martin Scorsese, Ridley Scott, Peter Jackson, and David Fincher. Conversely, the single English-language film with the highest number of votes, *The Shawshank Redemption*, has about 2.7 million votes, which is about 0.32% of all English votes.

Though these are considerably large percentages, other languages appear to be far more likely to have dominating figures and films. Hayao Miyazaki, the Japanese animator and director behind Studio Ghibli, accounts for 22.3% of all Japanese-language votes, almost 13 magnitudes greater than Christopher Nolan's proportion of English votes.

Parasite alone, through its international success, makes up 13.8% of all Korean-language votes. Bong Joon Ho, the director of *Parasite*, takes over 26% of the total share of Korean votes.

This appears in multiple languages, not just East Asian languages. For instance, when looking at the Danish language, 26% of all ratings come from the Academy Award winning director Thomas Vinterberg’s Danish films, such as *The Hunt* (2012) and *Another Round* (2020).

This poses interesting questions about which films and individuals outside of Hollywood are able to gain traction. This phenomenon does not exist for all languages, but it is evident that the likelihood of a single film, director, or actor dominating a language’s film market is possible in non-English languages. At this point, however, it does not seem plausible for an English film to amass somewhere around 16 million votes.

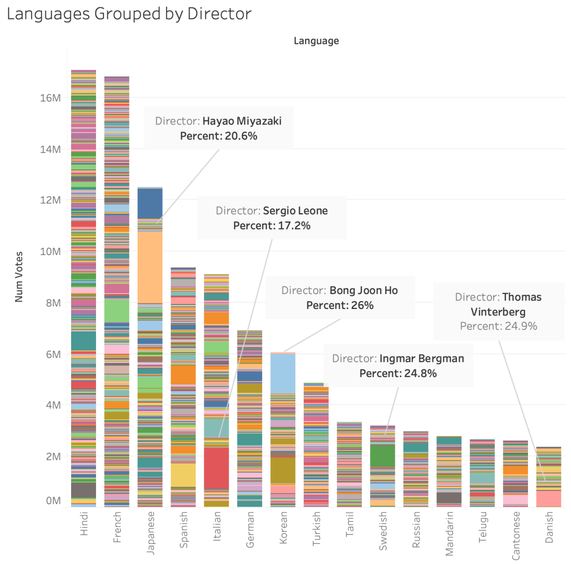


Figure 11: Languages grouped by director

6.6 Growth of Foreign Language Films

Across the past century, the number of English films being produced has grown exponentially. Along with this, the number of languages films are being produced in is also growing, however it is doing so at a linear rate. The increase in film production can easily be attributed to the growth of at-home entertainment, the ease of use and cost effectiveness of digital cameras, and the international market, particularly from Hollywood films. However, of the roughly 6,500 languages in the world today, in 2022 movies were produced in only 107 unique languages. For comparison, in 1952, there were only 33 languages movies were produced in. It is not clear to us what would have caused the number of languages for films to grow in this way, instead of increasing at the same rate as the movies being produced.

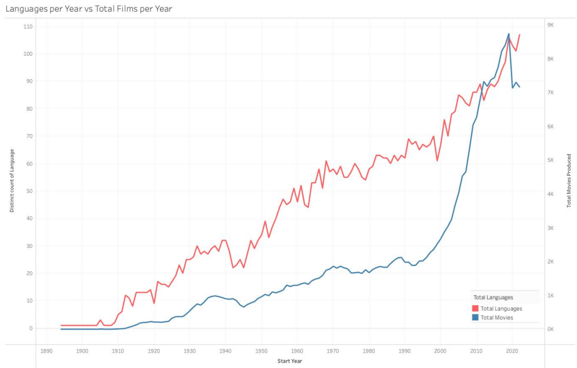


Figure 12: Number of films produced vs number of languages films were produced in

6.7 Number of Votes and the Top 250

IMDb famously has the “IMDb Top 250” which is a list of the 250 films with the highest ratings amongst regular voters. Many of the films in the list are amongst the most popular in the world, with the top three spots going to *The Shawshank Redemption*, *The Godfather* (Coppola, 1972), and *The Dark Knight* (Nolan, 2008). In the top ten of this, every film has over 500,000 votes. In total, there are 259 movies that have over 500,000 votes. However, just ten of these are non-English, yet all but one are in the IMDb Top 250. It’s also important to note that the one film not included in the Top 250 is *The Kashmir Files* (Agnihotri, 2022), which IMDb has purposefully excluded from the list due to suspected voter manipulation tactics. Of these nine films, they cross six languages: French, Hindi, Italian, Korean, Japanese, Portuguese, and Spanish. The titles are *The Good, the Bad, and the Ugly* (Leone, 1966), *Life Is Beautiful* (Benigni, 1997), *Amélie* (Jeunet, 2001), *Oldboy* (Park, 2003), *Parasite* (Bong, 2019), *Spirited Away* (Miyazaki, 2001), *City of God* (Meirelles, 2002) and *Pan’s Labyrinth* (Del Toro, 2006).

At the time of writing, 78% of films in the Top 250 are English-language. Of the thirty films in the Top 250 with fewest number of votes, eighteen of them are non-English, 60% of the bottom thirty, despite making up just 22% of films on their highly respected list.

6.8 Years

Over the course of the past century, there have been surprisingly few deviations between English and non-English films, whether you're looking at genre, ratings, or number of votes. In fact, today films are being made in other languages at the same rate as English films.

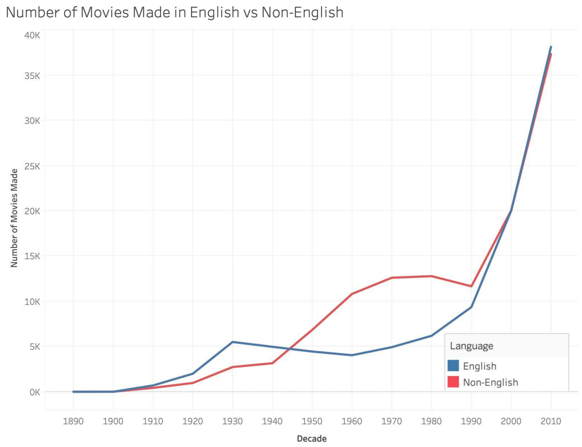


Figure 13: Total number of movies being produced per decade

As Figure 13 shows, between the 1950s and 1990s there were more films being produced in other languages, primarily European languages. Beyond these findings, there have been few differences over the years.

7 Applications

The overall question explored in this project is whether certain conditions lead to greater user engagement with non-English films. As previously articulated, this kind of engagement can be treated as a measure of success that is different from more traditional benchmarks such as box office proceeds or streaming views. We also believe that a platform like IMDb can lead to users discovering film and television titles they might not otherwise come across. With all of this in mind, our findings here can be applied in a number of different ways.

Perhaps the most obvious application is related to the potential investment in non-English film distribution in either the U.S. market or more globally. Our research revealed that while still behind English films overall, there is certainly a market for non-English films among IMDb users. Given a set amount of resources, our findings may help to decide which films should be prioritized.

Our findings support some expected results—a film by an established director, for example, has a high chance of user engagement. Taking this a step further, we suggest that based on our results any director that makes it into the IMDb Top 250 with a non-English film has an excellent chance of success with future films.

Another applicable finding relates to the language itself. We found that some languages appear to have an established market and therefore are a relatively safer bet: Spanish, French, German, Italian, and Japanese, for example. This may be due in part to the number of speakers globally for each of the aforementioned languages, but our research indicates that it is not wholly responsible for their success.

Perhaps more unexpectedly, given our findings when it came to user votes, genres, and "underrated gems," Indian language films may be a category with wide potential that has not yet been fully realized. The recent relative commercial success of Telugu film *RRR* (Rajamouli, 2022) could be seen as a more concrete example of this.

Some genres may also be a more attractive investment choice when it comes to non-English films. Drama in particular has had obvious and sustained success, while Documentary has seen a rise in popularity over the past three decades. Our research also suggested that films involving the genre combination of Crime, Drama, and Mystery have also done well.

A second potential application involves marketing, specifically marketing on the IMDb platform. We believe that our findings may help non-English films in choosing conditions that will encourage user engagement. If you have a film involving crime or mystery, for example, it may be beneficial to categorize the genre as Crime/Drama/Mystery. Additionally, if there are multiple languages spoken in the film it is certainly a positive to list all of them, rather than the main language.

Finally, our research could be applied with regards to IMDb users looking to discover new films. The underrated gems on our list provide non-English films that are highly rated with enough user votes (according to our definition) to offset bias, but are not globally well-known. IMDb's capability of introducing users to films they might not otherwise discover has certainly helped with its longevity, and we believe that our findings may expand on this.

ACKNOWLEDGMENTS

This project is a part of the CSPB 4502 Data Mining Course for Spring, 2023 taught by Professor Kristy Peterson.

REFERENCES

[1] Anon. 2023. IMDb Dataset - title.akas.tsv.gz. (March 2023). Retrieved March 18, 2023 from https://www.imdb.com/interfaces/

[2] Anon. 2023. IMDb Dataset - title.basics.tsv.gz. (March 2023). Retrieved March 18, 2023 from https://www.imdb.com/interfaces/

[3] Anon. 2023. IMDb Dataset - title.ratings.tsv.gz. (March 2023). Retrieved March 18, 2023 from https://www.imdb.com/interfaces/

[4] jasonlei0420. 2020. DS5230 movie recommendation system. (April 2020). Retrieved March 18, 2023 from https://www.kaggle.com/code/jasonlei0420/ds5230-movie-recommendation-system

[5] Jiawei Han, Micheline Kamber, and Jian Pei. 2012. *Data Mining Concepts and Techniques*, Burlington, MA: Elsevier.

[6] Keswanirohit. 2021. Netflix visualization and Eda. (February 2021). Retrieved March 18, 2023 from https://www.kaggle.com/code/keswanirohit/netflix-visualization-and-eda

[7] Khomani. 2022. Chantal Akerman first woman to top Sight and Sound's greatest all-time films poll. *Guardian*.

[8] Slayomer. 2020. Eda on IMBB film dataset. (April 2020). Retrieved March 18, 2023 from https://www.kaggle.com/code/slayomer/eda-on-imbb-film-dataset

[9] Stieg Larsson stats: By the numbers " in the bookroom: 2011. *https://web.archive.org/web/20110605085114/http://blog.libraryjournal.com/inthebookroom/2011/06/03/stieg-larsson-stats-by-the-numbers/*. Accessed: 2023-04-30.

[10] What are the top 200 most spoken languages?: *https://www.ethnologue.com/insights/ethnologue200/*. Accessed: 2023-05-03.