Non-English Film Engagement on IMDb

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

ABSTRACT
As media is increasingly globalized, the success in recent years of films like *Parasite* and *All Quiet on the Western Front* demonstrate the widespread potential for Non-English language films in the English-speaking market. 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 aim to explore user engagement with non-English language films listed on IMDb. We began by preprocessing and integrating several different datasets available from IMDb, pulling out the film titles and relevant attributes. From there, we collected language information for each title in order to assess the difference between English and Non-English films. We are currently in the process of applying data mining tools and methods to determine if there are certain conditions such as genre, language, or release year that lead to greater user engagement. We hope that this research will lead to interesting patterns and shed new light on this topic.

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; User engagement; Data visualization

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, 3 pages.*

1 Problem Statement and Motivation

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 number of ratings. We believe 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.

We have developed six key research questions for this project:

- Do non-English films have fewer reviews and ratings, regardless of their average rating?
- With increased globalization, are films being produced in more languages or fewer languages?
- Do certain genres of movies do better (better could be ratings or gross/net profit) when released at certain times of the year?
- Do certain languages do better when it comes to ratings/reviews overall?
- Do certain languages produce more films in certain genres and, if so, how does that affect the user engagement of a film?
- Does the number of language speakers impact the average number of ratings per film for that language?
- Do international films have higher ratings due to fewer view counts?
- Which films are “underrated gems,” that have high ratings, but relatively low

view counts?

We hope that our research can find patterns within non-English film engagement that can shed light on the willingness of users to watch content from countries outside of their own. English films became a dominant force in international markets since the Second World War 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.

2 Literature Survey

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, which 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. It is worth noting that some of the exploratory analyses may prove useful once we have applied our models in that we can compare the engagement patterns in non-English Language films with engagement patterns over all films.

3 Proposed Work

Firstly, the IMDb dataset will be organized. Cleaning unnecessary rows and columns to research for this project, and blank entries. For example, the data will be removed from the television cast and n/a data, only focusing on movies. Combining multi datasets to enable finding user engagement with non-English language films listed on IMDb. If a movie has multi-languages, we will trim it to one language, which is the origin country to be made.

Second, looking at data more carefully to check outliers. For example, how to consider how much minimum votes can be equal enough to treat against high votes movie using histograms. Exploring the organized dataset finds interesting patterns and relationships among attributes. For example, the data will have some linear regression correlation, similar group results by clustering, or nothing to relate some attributes by correlation matrix.

In conclusion, the explored patterns are to be displayed as graphs, and charts. Perhaps we can be making some storytelling about why this happened, if the results have some background we can think. The result is also evaluated on how accuracy leads to the result of the research. For example, we can examine the accuracy by measuring ROC/f1 score.

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 currently suspect that we will be using the following data sets from IMDb: title.basics.tsv.gz, title.ratings.tsv.gz, and title.akas.tsv.gz. The other available data sets relate to cast and crew information, as well as episode details for TV shows, which are outside the scope of this 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), an average ratings, and the number of unique votes.

Our second data set, Basics, has 9 attributes,

including tConst. Using this data set, we will be able to start the data reduction process using the following attributes: isAdult (a boolean to mark if it's an adult or non-adult title) which we will use to remove all adult content, titleType (such as TV show, episode, film, short film, etc.) which we will use 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 might help us catch errors. When we move to modeling, we will use this data set to access genres of films, the year the film was released, and the runtime.

Lastly, the third data set we will access is the akas (also known as) data set. This will tell us the language of the film, the regions in which the original title is used, as well as allowing us to access both the original title and the translated title.

5 Evaluation Methods

There are four main steps that we plan to do to evaluate our results. They are statistical analysis, accuracy checks, data visualization, and critical evaluation.

5.1 Statistical Analysis

For statistical analysis we will utilize clustering, regression, pattern mining, as well as other potential methods as we learn more techniques in class and become more familiar with the dataset.

5.2 Accuracy Checks

For accuracy checks we will develop test cases and comparison sets to ensure accurate data processing. We will also utilize a couple Python packages to fill in any data that may be missing from the dataset.

5.3 Data Visualization

For data visualization we will utilize graphs, plots, and other visual tools to evaluate and recognize patterns that may be found in our dataset.

5.4 Critical Evaluation

For critical evaluation we will scrutinize results for misleading strong association rules.

6 Tools

6.1 Github

We will be using Github repository to store and keep track of all of our project milestones and class assignments related to the project.

6.2 Discord

We will be using Discord for daily team communication and discussions. As well as relaying any blocks that may occur and how to effectively resolve issues.

6.3 Zoom

We will be using Zoom for weekly team meetings to more effectively communicate what each of us has been working on, as well as what should be worked on to accomplish our milestones.

6.4 Google Docs and Slides

We will be using Google docs and slides for related class project assignments, as this effectively allows multiple people to edit at the same time.

6.5 Python

We will be using the Python programming language to find trends/patterns related to the problems and questions we have for our dataset. Within Python we will be using Numpy and Pandas to help us navigate through and run calculations as needed in our dataset. We will also potentially be using Matplotlib to help visualize any discoveries made.

We will also be using Cinemagoer and TheMovieDB (TMDB). Both are Python packages that are able to retrieve/read in movie data and information from IMDb. They will help to fill in any potential missing data, as well as

verify that the data we have is correct as necessary.

6.6 Tableau

We will be using Tableau to help us find and visualize any simple and interesting patterns related to our dataset.

7 Milestones

7.1 Milestones

We have identified 5 key milestones for conducting our data analysis.

1. Data cleaning, integration, and reduction by 31/3/23 - We will have cleaned and processed the data to prepare it for analysis. This includes tasks such as removing duplicates, eliminating all non-feature films, handling missing values, and transforming data into a consistent format.
2. Search for patterns by 7/4/23 - We will have analyzed the data to identify patterns and relationships between variables.
3. Develop and test models by 21/4/23 - We will have built models to predict outcomes or explain relationships in the data and tested these models to ensure they are accurate and reliable.
4. Apply models to data set by 28/4/23 - We will have used models to make predictions and draw conclusions about the IMDb data.
5. Visualizations by 28/4/23 - We will have created visualizations to present the findings of the analysis in a clear and meaningful way. These visualizations will be able to help stakeholders understand the results of the analysis.

7.2 Milestones Completed

Data preprocessing has been completed. Though it did take us longer to complete than expected, we have successfully cleaned and reduced our data in a way that makes modeling and visualizing the data much simpler. We believe we have chosen the dataset that is most relevant to our research questions. We were able to use a Python library and API that allowed us to gather more information. In the end, we were able to obtain more information about movie languages.

We successfully researched and found answers to the correlations and patterns we were looking for in our research questions outlined in section 1. Because the data-preprocessing took longer than expected, this deadline also got postponed, but only by a few days.

7.3 Milestones To-Do

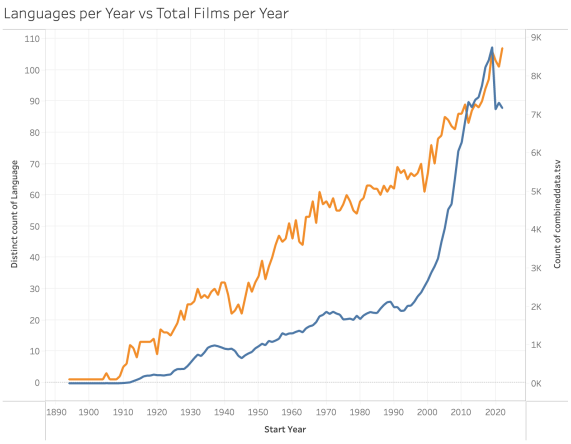
Currently, we are working on building models that will predict outcomes or explain relationships in the data. By the deadline, we will have tested these models to ensure they are accurate and reliable.

By our deadline of the 28th of April, we will have used models to make predictions and draw conclusions about the IMDb data. This deadline remains the same as originally planned.

We will have created visualizations to present the findings of the analysis in a clear and meaningful way. These visualizations will be able to help stakeholders understand the results of the analysis. This deadline remains the same. Our goal was to have this done by the 28th of April, however, we have already successfully created many useful visualizations that we can apply to our final report. We will leave this last milestone as to-do since there is still room for improvement and new discoveries.

8 Current Results

One result we have found is that the number of movies being produced started growing exponentially in the 2000s, which was the result of digital cameras, high cinema attendance, and the growth of VOD and home-entertainment accessibility. However, the number of languages films were produced in did not grow at the same rate, instead that grew only linearly.



Also we found a result whether non-English films have fewer reviews and ratings, regardless of their average rating or not. Running a T-test, the Non-English group of movies have significantly less votes (ratings). On average the Non-English films had 7131 fewer votes than English films (t_value = 46.068, $p < 0.001$). This assumes that variance is roughly equivalent, which was not tested for. Also, number of samples in each group Running an ANOVA for average rating by languages, there are significant differences between the averages. ($F = 42.84$, $p<0.01$)

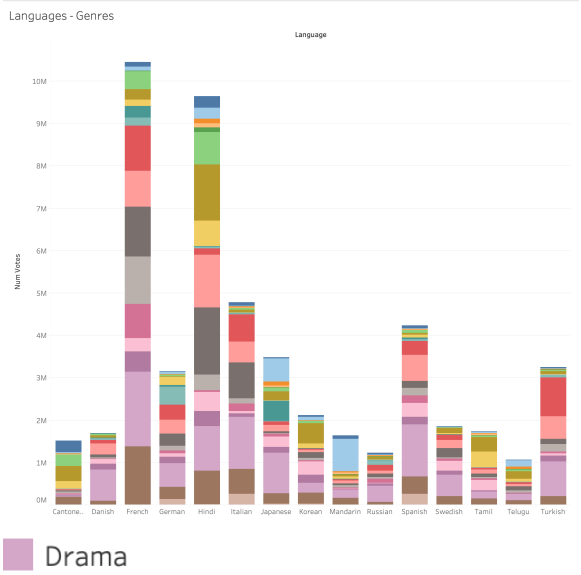
OLS Regression Results

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

In examining the user engagement with regards to genre, we have 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.

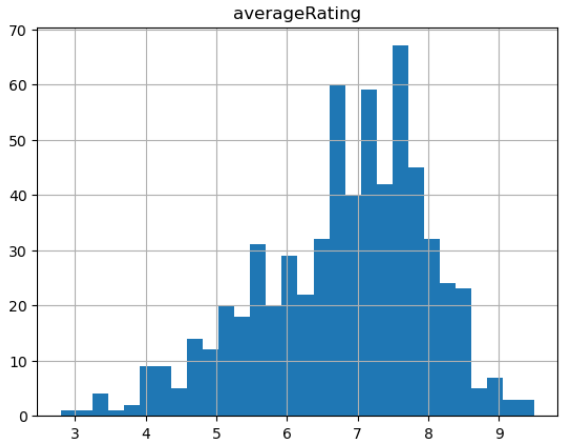


Drama films comprised over 22% of Non-English films in the dataset, and accounted for nearly 8.6% of all Non-English film user votes. By comparison, the Drama genre accounted for only 11% of English film titles and 3.2% of overall user votes. Additionally, of the user votes on English language Drama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

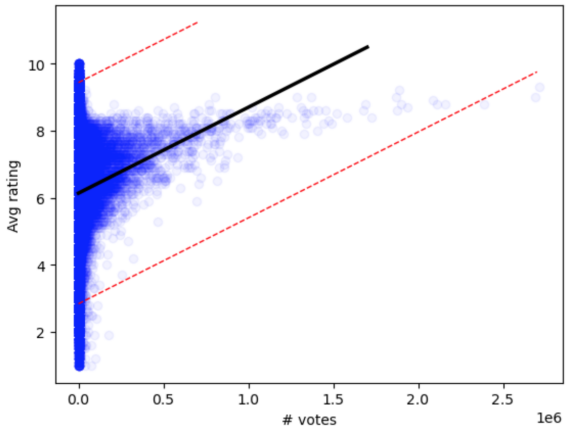
This finding suggests that Non-English Drama films have a higher chance of succeeding when it comes to user engagement over other genres, perhaps because Drama represents a much smaller portion of English language films and there is less competition. However, more work exploring the relationship between genre, user votes, user ratings, and individual language popularity still needs to be done before any definitive conclusions can be drawn.

While further examining movies based on individual languages, it was discovered that

movies in Bengali seemed to have a higher average rating, overall, when compared to all of the other languages.



Lastly, we tried to find out which films are "underrated gems," that have high ratings, but relatively low view counts. First, we examined the correlations between the number of votes and the average rating. The average rating distribution is similarly distributed in the chart between the average rating and the number of votes. Then, 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, which is about 1% confidence. These are defined as movies that are rated an average of 2.5 standard deviations above the average when controlling for the number of votes. This is not normal data, however. Perhaps standard linear regression is not an ideal method, but maybe Gamma regression is.



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] Slayomer. 2020. Eda on IMBB film dataset. (April 2020). Retrieved March 18,

2023 from https://www.kaggle.com/code/slayomer/eda-on-imbb-film-datase