Behind the Stars: Uncovering Hidden Adjustments in Letterboxd Film Ratings

Caio Santana Trigueiro
caiosantana@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Lucas Dayrell
lucasdayrell@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Arthur Buzelin
arthurbuzelin@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Guilherme H. G. Evangelista
guilherme.evangelista@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Caio Souza Grossi
caio.grossi@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Virgilio A. F. de Almeida
virgilio@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Wagner Meira Jr.
meira@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

ABSTRACT

Letterboxd's movie ratings influence millions, yet its scoring algorithm is opaque. We investigate the discrepancy between the platform's displayed score and the true user rating average, which we define as Δ. Analyzing a corpus of 1,737 Brazilian films and over 1.3 million ratings, we uncover the factors driving this distortion. Our analysis reveals a systematic algorithmic compression that pulls extreme scores toward the mean, with a strong negative correlation (-0.903) between a film's true rating and its Δ. Using K-Means, we identify four distinct rating distribution profiles (e.g., *Polarized*, *Highly-Rated*) and demonstrate that these profiles, along with genre, are significant predictors of the score adjustment. Niche genres like documentaries and musicals, which often exhibit polarized or extremely high ratings, are penalized most heavily. Furthermore, we find that popularity acts as a stabilizer; as a film's rating count increases, the magnitude of Δ decreases. Taken together, these results indicate that Letterboxd employs a normalization mechanism that mitigates the influence of outlier patterns, potentially fostering more representative aggregate scores and enhancing comparability across films. This study proposes greater transparency in these algorithms that shape cultural consumption.

KEYWORDS

Letterboxd, rating systems, algorithmic transparency, film recommendation, user behavior, score normalization, social platforms

1 INTRODUCTION

Television and film have been a core part of everyday life for decades. From movies to TV shows and soap operas, these forms of entertainment are part of what many people watch daily. In the past, deciding whether a movie was worth watching often came down to word of mouth recommendations or reviews in the news.

But with the rise of the Internet, that dynamic has changed. Today, people often turn to the electronic word-of-mouth (eWOM) on online platforms to decide what to watch, a practice that is particularly crucial for experiential products like movies [6]. Sites like Rotten Tomatoes and IMDb have long been popular, but in recent years, no platform has captured the attention of film fans quite like Letterboxd. [1,2]

Unlike older aggregators that foreground critic or professional reviews, Letterboxd centers popular opinion and social interaction. It highlights what friends are watching, liking, and discussing, turning reviewing into a community activity. As a result, titles tend to accumulate far more user ratings than on legacy sites. In our Brazilian-film snapshot, for example, City of God (2002)—the most-reviewed title in our dataset—shows roughly 110k logged reviews on Letterboxd versus only 1k on IMDb as of April 2025. This scale has made Letterboxd a hub for discussions spanning mainstream blockbusters to niche indie films and TV series.

In a time when public opinion carries more weight than ever, it's important to understand how platforms like Letterboxd shape our views. On the surface, Letterboxd appears straightforward: users can rate films from 0 to 5 stars, in half-star increments. These ratings are then aggregated into an overall score for each movie. **However, there's a catch: the displayed final score doesn't always match the ratings average.**

Letterboxd claims there was a recent update in their weighting algorithm for a "*better, fairer approach to weighting across the board, to more accurately reflect the Letterboxd community's global consensus for each film*" [3], but don't disclose how this score is calculated.

This discrepancy raises important questions. Given that a simple numerical average can be an unreliable proxy for true user sentiment [9], Letterboxd's weighting algorithm may represent an attempt to derive a more community centered score. With that in mind, it is essential to understand how the algorithm rewards or

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

[1] https://www.independent.co.uk/arts-entertainment/films/features/letterboxd-app-movies-celebrities-users-b2775948.html
[2] https://variety.com/vip/letterboxd-year-end-report-growth-1236277320/
[3] https://letterboxd.com/journal/the-score-new-weighted-average-ratings/

penalizes films rating patterns. Therefore, we propose the following research questions:

RQ1: What measurable patterns explain the discrepancy between a film's true rating and its displayed rating on Letterboxd?
RQ2: What are the broader implications of these patterns for film visibility, genre reception, and audience perception?

To explore this issue, we conducted a large-scale analysis of approximately 1,700 Brazilian films released between 1925 and 2025, collecting over 1.3 million user ratings through our own data scraping pipeline. We then explored two key metrics: the average rating actually given by users (*true rating*) and the official rating displayed by Letterboxd. The difference between these two values, denoted as Δ, became the central focus of our investigation.

We applied a variety of data mining techniques to understand what drives this discrepancy. First, we used linear regression to quantify how much of Δ could be explained by the true average alone. Then, we examined the role of popularity (measured by view counts) and analyzed how the variance in Δ changes as a film gains visibility. We also clustered film genres into macro-categories to assess whether certain types of content are more affected by the adjustment process. Additionally, we identified four typical patterns of rating distributions (e.g., polarized, consistently high-rated, balanced), and studied how these profiles relate to the score distortion.

Finally, we created profiles of similar genres and distribution of ratings to prove the relationship between them through statistical association tests. Our results reveal that Letterboxd's algorithm systematically compresses both extremely high and low scores, especially in nich genres, as well as more polarizing genres such as documentaries and musicals. Meanwhile, mainstream productions exhibit more neutral or slightly downward-adjusted scores. This study offers a data-driven perspective on how the platform's opaque scoring mechanism shapes public perception, and it underscores the importance of transparency in systems that influence collective cultural taste.

2 RELATED WORK

The literature on rating and recommendation systems—both for films and for other cultural platforms—points to three main research streams that matter for this study: (i) **scoring-algorithm modeling**, (ii) **social-influence and popularity effects**, and (iii) **structural analysis of preference data**.

Below, we summarize the key findings of each stream and explain how they connect to our investigation of Letterboxd ratings.

2.1 Scoring-Algorithm Modeling

Lu *et al.* [8] propose a Bayesian mechanism to rank TV series on IMDb, showing that probabilistic approaches can soften extreme ratings in dense datasets. Although we do not uncover Letterboxd's exact formula, we present empirical evidence of a similar *shrinkage* effect. Factorization-based recommenders (SVD) have been applied directly to Letterboxd data [2], revealing how observation bias hurts predictive accuracy. D'Addio *et al.* [5] explore textual-feature extraction from reviews, reinforcing the value of latent attributes beyond raw scores, which is a direction we flag for future work.

2.2 Social Influence and Popularity

Because review platforms are social, popularity, conformity, and time dynamics shape how people perceive item quality. Yang *et al.* [11] show that "outlier" reviews on Amazon can be judged more helpful, pointing to a social selection process. Yeste and Calduch-Losa [12] analyze Twitter hashtags to forecast movie-opening success, while Crisci *et al.* [4] use Twitter metrics to predict TV audiences. Our finding that the variance of the discrepancy Δ falls as *views* grow aligns with the idea that popularity acts as a statistical stabilizer.

2.3 Preference Structure and Clustering

Discovering hidden consumption patterns is another recurring theme. Ridenour and Jeong [10] cluster Goodreads books by co-reading, an approach similar to our genre and rating-distribution clustering. Chen and Dai [3] mine topics and social signals to link box-office results with online chatter, showing that metadata and social media signals can explain performance swings. Acerbi *et al.* [1] apply sentiment analysis to Twitter reactions to the documentary *Our Planet*, confirming that niche content—such as documentaries and musicals—tends to attract more polarized ratings, a pattern that also appears in our cross-analysis of genre and score profile.

2.4 Synthesis

Taken together, these studies suggest that (a) algorithms can invisibly impose statistical corrections on average ratings; (b) popularity and social influence modulate those corrections; and (c) latent preference structures are essential to explain observed discrepancies. Our work adds to this body by quantifying these factors at a national scale and by calling out the lack of transparency in Letterboxd's displayed rating, filling a niche that the literature has so far left open.

3 DATASET

To investigate the discrepancy between displayed and actual average ratings on Letterboxd, we constructed a custom dataset of 1,737 Brazilian films released between 1925 and early 2025, as illustrated in 2. Data collection took place during the first week of April 2025 (exactly one month after the 2025 Oscars) which provided a stable snapshot of recent user activity on the platform.

We chose to focus our analysis on Brazilian films for both practical and conceptual reasons. On a personal level, the authors are based in Brazil, which makes the national cinema context more familiar and relevant to our research interests. From an analytical standpoint, Brazilian cinema offers a compelling balance: it is diverse and culturally rich enough to reveal meaningful patterns in audience behavior and genre reception, yet relatively small compared to global film industries such as Hollywood. This allows for a comprehensive, nearly exhaustive collection of films without the need for sampling or aggressive filtering. By targeting a complete snapshot of Brazil's filmography, we ensure that our analysis captures a wide spectrum of production types, genres, and audience dynamics, while remaining feasible within the time and computational resources available.

Before exploring the specifics of our dataset and data collection process, it is worth briefly examining the temporal distribution

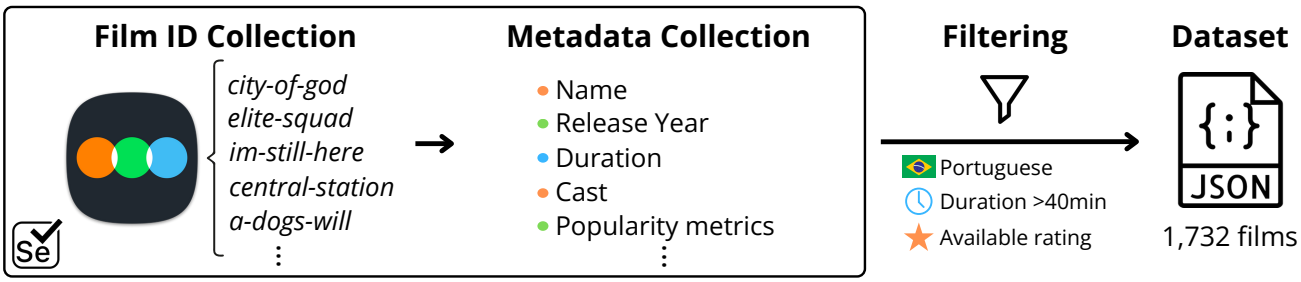

Figure 1: Diagram summarizing the data collection pipeline.

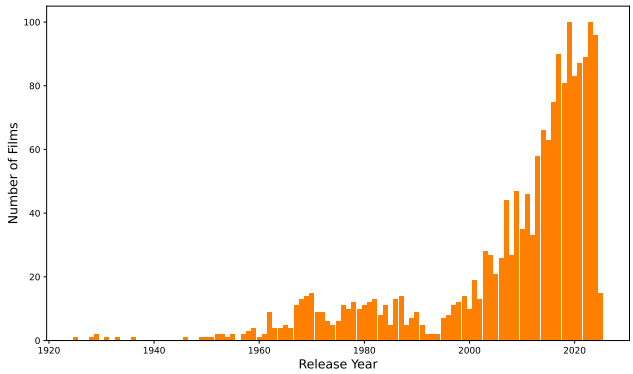

Figure 2: Number of films by release year, ranging from 1925 to early 2025.

of film releases in our sample as it gives interesting insight on Brazilian cinema evolution. As shown in Figure 2, the number of Brazilian feature films released annually has grown substantially over the past two decades, with notable acceleration from the early 2000s onward. This upward trend aligns with the implementation of key public policies and funding programs aimed at revitalizing the national film industry, such as the Audiovisual Law (Lei do Audiovisual, 1993), the National Cinema Agency's (ANCINE) incentive mechanisms established in the early 2000s, and the Fundo Setorial do Audiovisual (FSA) launched in 2006. Increased state-backed investment, combined with regional production funds and co-production agreements, expanded the capacity of Brazilian studios and facilitated a more diverse range of productions and creating a richer and more varied domestic catalog.

3.1 Data Collection Pipeline

The dataset was created via a two-stage scraping pipeline built with Selenium. In the first stage, we accessed Letterboxd's country-specific browsing interface, which lists approximately 20,000 titles under the Brazil category. However, this list includes any film shot in Brazil, not exclusively Brazilian productions. Therefore, we collected the unique slugs (film IDs) sorted by popularity and performed post-hoc filtering to isolate genuinely Brazilian titles.

In the second stage, we programmatically visited each film's page to extract metadata from multiple subpages (e.g., '/genres', '/details', '/reviews'). The extracted attributes included are shown on the example in Table 1.

Table 1: Summary of Movie Metadata: *City of God*

Field	Value
Name	City of God
Original Name	Cidade de Deus
Release Year	2002
Language	Portuguese
Duration (min)	129
Directors	Fernando Meirelles
Genres	Drama, Crime
Cast	Alexandre Rodrigues, Leandro Firmino...
Letterboxd Rating Avg	4.54
True Rating Avg	4.52
Reviews Count	109,689
Rated Count	752,851
Views Count	1,031,915
Listed Count	204,979
Likes Count	418,448
Fans Count	48,629
Star Ratings Distribution	570, 1019, 755, 3686, 4042...

Fallback strategies and retry logic were employed to ensure robust extraction even in the presence of incomplete or delayed HTML loading. All collected data was stored in CSV format and later converted to a structured JSON format for downstream analysis.

3.2 Filtering and Preprocessing

After collection, the dataset was filtered using a three-step process:

- **Language**: Only films with Portuguese as the primary language were retained.
- **Duration**: We filtered out short films and retained only *feature-length* films, following the Academy's (Oscars) definition of features as works with runtime over 40 minutes[4].
- **Rating Availability**: Films with fewer than 200 ratings were excluded, since Letterboxd does not compute or display an official rating (*Letterboxd Rating*) below this threshold.

After applying these filters, the final dataset contained 1,737 films. Each entry was converted into a JSON object grouped under

[4]https://www.oscars.org/sites/oscars/files/95aa_feature_film.pdf

the key 'movies'. This allowed for easier retrieval and aggregation across different dimensions (e.g., by genre or distribution profile). Figure 1 summarizes sections 3.1 and 3.2, illustrating our pipeline.

3.3 Ethical Considerations

All data collected was publicly available on Letterboxd. We did not access any user-identifiable or private information, and no data was obtained through authenticated sessions or third-party APIs. Due to the terms of use of the platform, the dataset is not redistributed, but the data collection pipeline and filtering scripts can be shared upon request to support reproducibility and peer review.

4 EXPERIMENTS

To provide a more intuitive narrative, we present our experiments in the chronological order in which our team explored and uncovered the underlying dynamics behind Letterboxd's rating system.

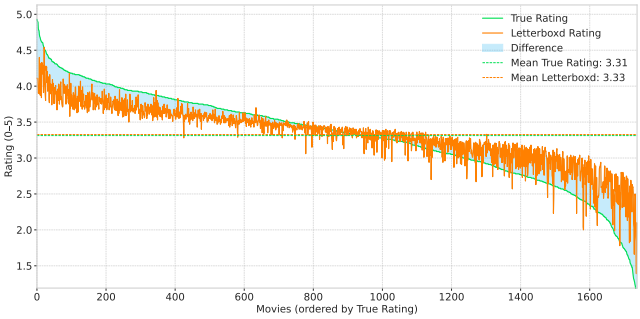

Figure 3: Difference between Letterboxd and True ratings (Δ), ordered by true rating.

4.1 Understanding the Real Grading

Our first and most immediate step was to quantify the discrepancy between the rating displayed by Letterboxd and the actual average rating given by users. For each film X, we define the following:

- $R_{\text{official}}(X)$: the official rating displayed on Letterboxd.
- $R_{\text{true}}(X)$: the true rating, computed as the arithmetic mean of all user ratings for that film.

We define the rating delta, $\Delta(X)$, as the difference between these two values:

$$\Delta(X) = R_{\text{official}}(X) - R_{\text{true}}(X) \tag{1}$$

This value captures how much the platform adjusts (or distorts) the visible score of a film compared to its actual average. A positive delta means the displayed score is higher than what users gave on average, while a negative delta indicates a downward adjustment.

Figure 3 illustrates the behavior of all 1,700 Brazilian films in our dataset, each represented by two lines: the average rating given by users (in green) and the official rating displayed by Letterboxd (in orange, with the difference highlighted in blue. Films are ordered by their true rating, from highest to lowest.

A few important patterns emerge. First, we observe a distribution reminiscent of a logistic distribution with very few films achieve extremely high or extremely low scores, while most cluster around the middle. This aligns with general expectations about large-scale user-generated content.

The core insight, however, lies in the gap between the two curves. Letterboxd's algorithm consistently pulls ratings toward the center. Movies with lower user ratings tend to receive a positive boost in their displayed score, while those with very high user ratings are slightly downgraded. This suggests the presence of a normalization or compression mechanism within the platform's scoring system.

This interpretation is further supported by Figure 4, which shows a strong negative Pearson correlation of -0.903 between the true rating and the delta. In other words, the higher a film's actual average rating, the more likely it is to be penalized by the Letterboxd algorithm, and conversely, the lower the true rating, the more likely it is to be boosted.

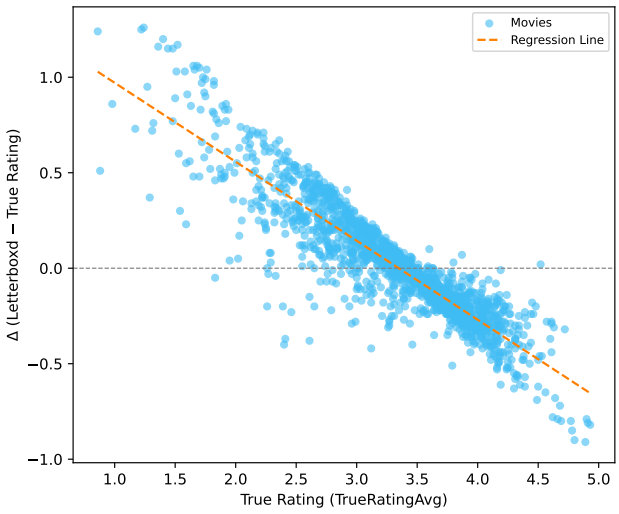

Figure 4: Correlation between Δ and True Rating.

Still, these visualizations only tells part of the story. While the general trend is clear, many outliers remain, with some films diverging significantly from the average behavior. These anomalies highlight that although centralization is a dominant trend, it does not act uniformly.

4.2 Movie Popularity

With the understanding that Letterboxd tends to adjust movie ratings toward a central value, we now investigate whether a film's popularity, measured by the number of users who rated it, correlates with the amount of interference applied to its score.

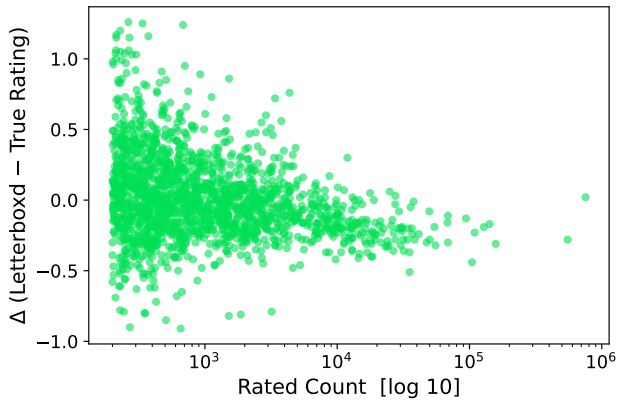

Figure 5: Relationship between the number of ratings a movie receives (log scale) and its rating delta (Δ). Each point represents a film.

As shown in Figure 5, there is a clear pattern: less popular movies tend to exhibit much greater variation in Δ, suggesting that their displayed ratings are more heavily adjusted by Letterboxd. In contrast, widely rated and well-known films show smaller deltas, implying that their scores are more stable and less manipulated.

Although the Pearson correlation coefficient is relatively small ($r = -0.083$), the relationship is statistically significant ($p = 0.000573$). This allows us to reject the null hypothesis and confirm that popularity, while not the strongest factor, does play a role in how much a film's score is altered by the platform.

4.3 Genre Clustering and Bias Patterns

Next, we turned our attention to the genres of the films, aiming for a more semantic perspective on rating distortion. Genre analysis could reveal whether certain types of content are more systematically affected by the platform's weighting algorithm. However, working directly with individual genres posed a challenge: the dataset includes nearly 20 distinct categories, and each film may be associated with multiple ones. To reduce dimensionality and reveal more general trends, we clustered the genres into broader *macro-genres* based on how frequently they co-occurred across films.

To accomplish this, we constructed a binary matrix where each row represented a genre and each column a film, with entries indicating whether a given genre was assigned to that film. Using this matrix, we computed the pairwise Jaccard distance between genres, which quantifies dissimilarity based on the sets of films each genre appears in. Formally, for two genres g_1 and g_2, the Jaccard distance is defined as:

$$d_J(g_1, g_2) = 1 - \frac{|F_{g_1} \cap F_{g_2}|}{|F_{g_1} \cup F_{g_2}|}$$

where F_{g_i} is the set of films tagged with genre g_i.

We then applied hierarchical agglomerative clustering using average linkage on the resulting distance matrix to group genres with similar co-occurrence patterns. We selected four clusters for interpretability, but excluded one of them from further analysis, as it contained only the genre *TV Movie*, which was predominantly associated with television content and represented less than 1% of the films.

The remaining clusters revealed three meaningful macro-genre groupings:

- **Cluster 1 - Narrative Fiction**: This cluster grouped 14 genres commonly found in fictional storytelling, including Drama, Comedy, Romance, Action, Horror, and Animation. It comprised the majority of the dataset, with over 1,400 films.
- **Cluster 2 - Historical/War**: A smaller but coherent group (5% of the dataset) composed solely of films labeled with the genres History and War. These films typically portray real or dramatized historical events and conflicts.
- **Cluster 3 - Musical/Documentary**: A distinct category focused on non-narrative or hybrid formats, containing Documentary and Music related films. These films often have a more intimate, factual, or performative nature and are less driven by conventional plots and represented about 400 of the 1737 total.

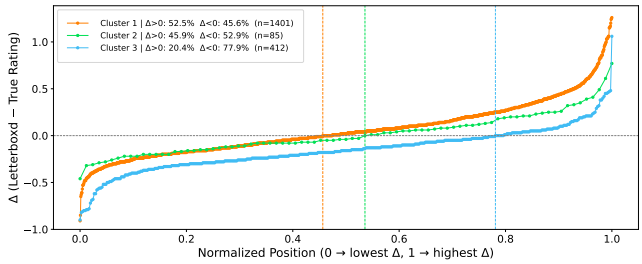

Figure 6: Normalized distribution of rating distortion Δ across macro-genres. The dashed line marks the threshold where Δ = 0, dividing overestimated and underestimated films.

These macro-genres provided a principled way to reduce the dimensionality of genre data while preserving semantic cohesion. In the following, we analyze how these groups differ in terms of their rating distortion Δ.

Figure 6 presents the normalized distribution of Δ = *Letterboxd Rating - True Rating* for each macro-genre, allowing for a direct comparison of distortion patterns across groups. Films are ordered by increasing Δ within each group, and the x-axis reflects their normalized rank. The dashed vertical line marks the point at which Δ = 0, dividing the films into two regimes: those *underestimated* by the platform (to the left) and those *overestimated* (to the right).

A clear distinction emerges across the curves. The **Musical/Documentary** cluster, in the blue line is visibly shifted to the left: around 78% of its films fall below the Δ = 0 line, indicating that these titles are systematically rated lower by the Letterboxd algorithm compared to their actual user averages. This suggests a consistent penalization effect on more niche or less mainstream genres.

In contrast, the **Narrative Fiction** cluster—in orange—has a more symmetric distribution centered near the dashed line, with

a slight tendency toward positive Δ. This implies that the platform either preserves or mildly boosts the ratings of these films. Approximately 52.5% of films in this group are overestimated.

The green **Historical/War** cluster is not the counter-pole to Musical/Documentary. Its Δ distribution concentrates near zero, with moderate dispersion and no pronounced directional skew. Given the relatively small number of titles in this group, estimates are noisier, but the pattern aligns more closely with the near-neutral behavior of Narrative Fiction than with any extreme shift.

Taken together, these results indicate that genre plays a *substantial* role in how Letterboxd adjusts a film's rating. The platform's algorithm appears to compress or normalize scores unevenly across different types of content, prompting us to investigate the underlying reasons for this behavior.

4.4 Rating Distribution Profiles

Beyond genre and popularity, the manner in which ratings are spread across the 0.5 to 5-star scale can offer valuable *insights* into a film's reception. The same average score can conceal very different realities: a tepid consensus, with many middling ratings, or a polarized reception, with a split between very high and very low scores. We hypothesized that Letterboxd's algorithm does not treat these distribution patterns identically.

Building on the previous section, we also considered that different result across genres might be naturally tied to distinct rating distribution patterns among users. These distributions are likewise directly related to the pure average rating—that we proved having a great impact in the Delta value— therefore understanding them is essential.

To investigate this, we employed clustering to segment films based on their voting patterns. Using the K-Means algorithm, we grouped the films into k=4 clusters (calculated with silhouette index), where each title was represented by a vector of its percentage rating distribution across 10 bins (from 0.5 to 5.0). The analysis, ilustrated in 7 revealed four distinct and interpretable rating distribution profiles. The first, termed **Highly-Rated**, is characterized by a high concentration of ratings at the top of the scale. A second pattern is the **Polarizing** profile,—with the lowest count, representing only 3% of the dataset—it exhibits an abnormally high 5 stars rating frequency but interestingly also an above average 0.5 (the lowest possible), typical of "love-it-or-hate-it", or review bombing victim films. In contrast, the **Tepid/Balanced** profile features a majority of ratings concentrated in the intermediate to low range, with lesser representation in the highest scores. Finally, the **Well-Distributed** profile shows a peak around 3.5 to 4.0 stars, suggesting a generally positive but not overwhelmingly acclaimed reception, representing about 42% of the dataset and being the most aligned with the global average.

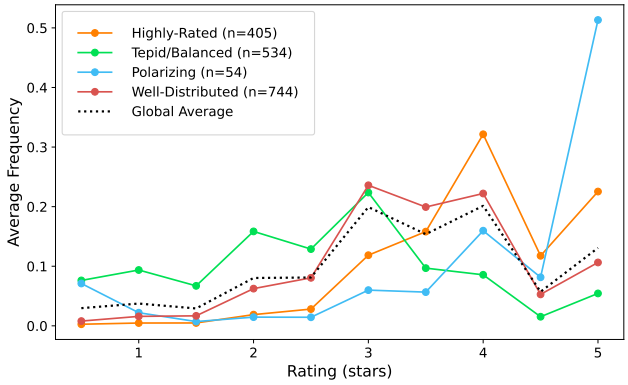

Figure 7: Visualization of the four average rating distribution profiles identified by the K-Means algorithm.

Upon analyzing the rating discrepancy (Δ) for each of these profiles (Figure 8), clear differences emerge. The **Highly-Rated** profile experiences the strongest downward adjustment, with 97.8% of its films showing $\Delta < 0$, indicating a systematic penalty on universally well-received titles. In contrast, the **Tepid/Balanced** profile is overwhelmingly boosted, with 91.9% of films having $\Delta > 0$, suggesting that the platform tends to elevate the scores of titles with middling consensus. The **Polarizing** profile shows a mixed pattern, probably due to the "love-it-or-hate-it" logic combined with it representing a smaller sample, but overall it leans toward negative adjustments (85.2% with $\Delta < 0$). Finally, the **Well-Distributed** profile appears closer to neutral, though a small majority (57.5%) still receive a positive adjustment. These results reinforce that Letterboxd's weighting mechanism interacts strongly with the underlying shape of the rating distribution, not just its average.

An additional phenomenon worth noting is the visible drop in frequency across all profiles right at the 4.5-star mark, compared to it's neighbors, breaking the almost normal distribution pattern in the global average. This abrupt decline is likely driven by users' natural tendency to round their ratings to the nearest half-star, particularly when uncertain between giving a film the maximum rating or not. Prior research has shown that people often round toward "cleaner" or more psychologically satisfying numbers, especially when presented with limited rating options [7].

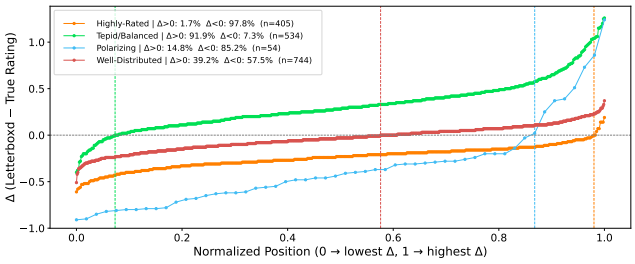

Figure 8: Density curves comparing the distribution of Δ for each of the four rating profiles.

4.5 Genre–Profile Association

To verify our hypothesis we finally explored whether the two variables were statistically related. Specifically, we aimed to test whether a movie's genre cluster could be predictive of its typical rating distribution profile.

To analyze this relationship, we quantify the association between macro-genres and rating–distribution profiles through a chi-square test applied to the contingency table between these two variables. The results indicate a statistically significant relationship ($\chi^2 = 199.81$, $df = 6$, $p = 2.08 \times 10^{-40}$), with a moderate effect size (Cramer's $V = 0.229$). To identify the strongest contributors to this association, Table 9 reports the *standardized residuals* (z) for each cell.

Positive (negative) values indicate genre–profile combinations that occur more (less) frequently than expected under the assumption of independence. Two main patterns emerge: (i) *Documentary/Musical* is strongly over-represented in the *Highly-Rated* and *Polarizing* profiles, and under-represented in *Tepid/Balanced*; (ii) *Narrative Fiction* shows the opposite trend, being over-represented in *Tepid/Balanced* and under-represented in the more extreme profiles. These deviations reinforce the notion that genre and rating–distribution shape are not independent, and help explain the uneven rating distortion Δ observed across genres.

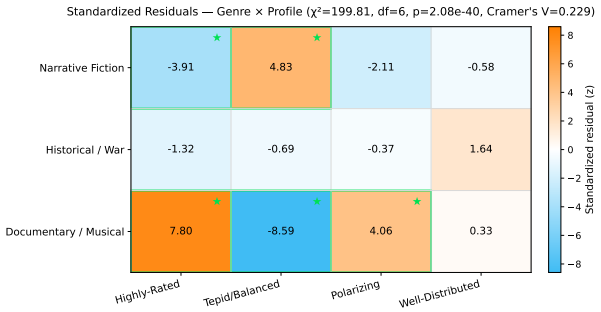

Figure 9: Standardized residuals from the chi-squared test for the association between *macro-genres* and rating distribution profiles. Positive values (orange) indicate over-representation, while negative values (blue) indicate under-representation. Cells outlined in green and marked with a star denote statistically significant deviations ($|z| \geq 3$).

These correlations confirm that genre and rating behavior are not independent. *Documentary/Musical* films are more likely to elicit polarized or extremely positive reactions from viewers—possibly due to their niche appeal or emotional content—while *Narrative Fiction* films tend to follow more uniform and central patterns. These genre-profile affinities may partly explain why certain genres experience stronger distortions in their displayed ratings, given Letterboxd's tendency to suppress extreme patterns in the interest of global consensus.

5 CONCLUSION

Across all experimental angles—average score distortion, popularity trends, genre effects, and distribution profiles—our results converge toward the same overarching insight: **Letterboxd's algorithm systematically penalizes deviations from normative patterns**. The platform appears to apply a normalization mechanism that reduces the influence of extreme user behavior, whether in the form of unusually high or low ratings, niche genres, or polarizing audience reception.

Among all experiments, the one that provided the clearest evidence of this behavior was the analysis of *rating distribution profiles*. We found that films with atypical voting patterns—such as highly polarized or universally acclaimed works—tend to be adjusted more heavily. These patterns often lead to either very high or very low average ratings, which are then moderated by the platform's final displayed score.

Interestingly, this distortion is not random. It aligns with underlying correlations between other factors. For example, *more popular films*, which attract broader and more diverse audiences, tend to have rating distributions that approximate the overall population average. These films naturally require less adjustment, and thus their Δ values are closer to zero.

Conversely, genres like *Documentary/Musical* show a strong connection to polarized or highly positive distribution profiles. This is likely due to the self-selecting nature of their audiences: viewers who engage with these films often already hold strong opinions or emotional investment in the subject matter—such as fans watching a concert documentary—which amplifies rating extremity and, consequently, the need for normalization.

Taken together, the evidence suggests that Letterboxd's weighting algorithm is not arbitrary. On the contrary, it appears to fulfill the platform's stated goal: "to more accurately reflect the Letterboxd community's global consensus for each film." By tempering the effects of rating extremes and prioritizing statistical centrality, the system seems to produce a more stable and representative measure of collective sentiment.

While this normalization comes at the cost of distorting certain niche or fan-driven works, the broader tradeoff may, in fact, support a fairer and more balanced discovery experience for the general user. However, this moderated score is the primary input for the platform's recommendation system, meaning surfaced suggestions are based on this moderated view of the community's consensus rather than on the raw user sentiment.

Appendix evidence. The ranked lists in the Appendix (Tables 2 and 3) illustrate this mechanism concretely. The *True Rating Top 10* is dominated by titles in the *Documentary/Musical* macro-genre (e.g., concert films and a club documentary) and *nearly all* of these entries exhibit large negative deltas ($\Delta < 0$), consistent with a strong downward adjustment of extremely high user averages. By contrast, the *Letterboxd Rating Top 10* contains more widely recognized, mainstream titles and shows deltas clustered closer to zero, aligning with our findings that popularity correlates with smaller adjustments and that the platform's weighting compresses extremes while preserving central, consensus-driven scores.

6 LIMITATIONS AND FUTURE WORK

While our analysis revealed consistent patterns in how Letterboxd adjusts film ratings, there are some limitations worth noting. Most notably, our study focused solely on quantitative data available

on film-level pages. We did not have access to certain potentially important metadata that could further illuminate the mechanics behind the algorithm's behavior. These include:

- The **timestamp of individual reviews**, which could help investigate temporal effects such as recency bias or the impact of review bursts.
- The **origin and profile of the reviewer**, including their geographic location, level of activity on the platform, or account age. These could help identify biases related to early adopters, regional reception, or even spam/bot behavior.

Although we believe these features might hold some explanatory power, we found the results obtained from the available metadata to be robust and insightful. The absence of these variables did not prevent us from uncovering meaningful and converging patterns in the platform's scoring mechanism.

Our study also opens the door for several directions in future research. First, we collected a rich set of additional attributes—such as film duration, director and cast lists, and user engagement metrics (e.g., number of fans, likes, or list appearances)—which could be leveraged to build predictive models or to analyze how artistic and social factors influence a film's reception.

Another promising avenue is the analysis of **textual reviews**. While not included in our current pipeline, these reviews could be incorporated in future crawls and analyzed using natural language processing (NLP) techniques. Sentiment analysis, topic modeling, and emotion detection could all add interpretability to the observed patterns, shedding light on the reasoning behind polarized scores or niche audience dynamics.

Finally, the framework we introduced can be extended beyond the Brazilian film corpus to compare national trends with international ones, or to study how algorithmic adjustments vary over time in response to platform policy changes.

REFERENCES

[1] Alberto Acerbi, John Burns, Unal Cabuk, Jakub Kryczka, Bethany Trapp, John Joseph Valletta, and Alex Mesoudi. 2023. Sentiment Analysis of the Twitter Response to Netflix's *Our Planet* Documentary. *Conservation Biology* 37, 4 (2023), e14060. https://doi.org/10.1111/cobi.14060

[2] Aditya Bhardwaj, Chirla Rushil Reddy, and Palak Arora. 2023. Movie Recommendation System Using SVD (Letterboxd). *International Journal of Advanced Research in Computer and Communication Engineering* 12, 10 (2023). https://doi.org/10.17148/IJARCCE.2023.121013

[3] Yinchang Chen and Zhe Dai. 2022. Mining of Movie Box Office and Movie Review Topics Using Social Network Big Data. *Frontiers in Psychology* 13 (2022), 903380. https://doi.org/10.3389/fpsyg.2022.903380

[4] Alfonso Crisci, Valentina Grasso, Paolo Nesi, Gianni Pantaleo, Irene Paoli, and Imad Zaza. 2018. Predicting TV Programme Audience by Using TwitterBased Metrics. *Multimedia Tools and Applications* 77, 10 (2018), 12203–12232. https://doi.org/10.1007/s11042-017-4880-x

[5] Rafael M. D'Addio, Marcos A. Domingues, and Marcelo G. Manzato. 2017. Exploiting Feature Extraction Techniques on Users' Reviews for Movies Recommendation. *Journal of the Brazilian Computer Society* 23 (2017), 7. https://doi.org/10.1186/s13173-017-0057-8

[6] Kacy Kim, Sukki Yoon, and Yung Kyun Choi. 2019. The effects of eWOM volume and valence on product sales – an empirical examination of the movie industry. *International Journal of Advertising* 38, 3 (2019), 471–488. https://doi.org/10.1080/02650487.2018.1535225 arXiv:https://doi.org/10.1080/02650487.2018.1535225

[7] Vassilis Kostakos. 2009. Is the Crowd's Wisdom Biased? A Quantitative Analysis of Three Online Communities. In *2009 International Conference on Computational Science and Engineering*. IEEE, 251–255. https://doi.org/10.1109/cse.2009.491

[8] BoYang Lu, Jia Li, YuZhong Chen, and Hao Xu. 2017. Evaluation of the Television Dramas Ranking Using the Bayes' Theorem. In *Proceedings of the 3rd Annual International Conference on Social Science and Contemporary Humanity Development (SSCHD 2017) (Advances in Social Science, Education and Humanities Research, Vol. 90)*. Atlantis Press, 155–157. https://doi.org/10.2991/sschd-17.2017.31

[9] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1909–1918. https://doi.org/10.1145/3178876.3186168

[10] Laura Ridenour and Wooseob Jeong. 2016. Leveraging the Power of Social Reading and Big Data: An Analysis of CoRead Clusters of Books on Goodreads. In *iConference 2016 Proceedings*. https://hdl.handle.net/2142/89314

[11] Kunhao Yang, Itsuki Fujisaki, and Kazuhiro Ueda. 2023. Social Influence Makes Outlier Opinions in Online Reviews Offer More Helpful Information. *Scientific Reports* 13 (2023), 9625. https://doi.org/10.1038/s41598-023-35953-4

[12] Víctor Yeste and Ángeles CalduchLosa. 2022. Exploratory Twitter Hashtag Analysis of Movie Premieres in the USA. In *Desafíos Audiovisuales de la Tecnología y los Contenidos en la Cultura Digital*. McGraw–Hill Interamericana de España S.L., 169–187.

A TABLES OF THE TOP 10 RATED MOVIES ON BOTH RATING TYPES

A.1 Letterboxd Rating top 10

Table 2: Top 10 films by Letterboxd rating, with Δ = Letterboxd − True

Rank	Title (Year)	Letterboxd	Δ
1	City of God (2002)	4.54	+0.02
2	O Auto da Compadecida (1999)	4.40	-0.32
3	I'm Still Here (2024)	4.33	-0.28
4	Twenty Years Later (1984)	4.32	-0.29
5	Central Station (1998)	4.32	-0.31
6	Playing (2007)	4.29	-0.28
7	Emicida: AmarElo – It's All for Yesterday (2020)	4.28	-0.20
8	Master, a Building in Copacabana (2002)	4.27	-0.18
9	Hilda Furacão (1998)	4.23	-0.37
10	Rio, Zona Norte (1957)	4.22	-0.24

A.2 True Rating Top 10

Table 3: Top 10 films by true rating, with Δ = Letterboxd − True

Rank	Title (Year)	True	Δ
1	Turnê Anti-Herói (Ao Vivo) (2020)	4.93	-0.82
2	Jão: SuperTurnê ao Vivo (2024)	4.91	-0.81
3	SuperTurnê: The First and Last Night (2025)	4.90	-0.79
4	Herobrine: The Legend (2013)	4.89	-0.91
5	Bittersweet Memories: Isolated for 14 Days to Create a Masterpiece (2021)	4.80	-0.90
6	Os Bagunceiros (2021)	4.78	-0.85
7	Todo Dia é 4 de Novembro: O Fluminense Conquista a América (2023)	4.77	-0.80
8	O Auto da Compadecida (1999)	4.72	-0.32
9	Validation: Isolated for 7 Days to Create an EP (2021)	4.69	-0.80
10	Paul McCartney: Got Back (2023)	4.68	-0.72