Data-Driven Multidimensional Evaluation of Cultural Works

Shuyang Yu

International Curriculum Center, The High School Affiliated to Renmin University of China, Beijing, China

Hua.yan7@icloud.com

Keywords: Cultural analytics, narrative modeling, fairness, success prediction

Abstract: This paper proposes a data-driven multidimensional framework to evaluate cultural works across films, television series, and novels. The framework integrates character relationship networks, narrative structure and perspective, audience emotion and empathy, and fairness-aware representation indices into four comprehensive measures: Representation Equity, Narrative Richness, Character Attractiveness, and Audience Engagement. We construct a real multi-source dataset spanning 2018–2023 and benchmark three predictive models for a normalized success score. Experiments show that gradient boosting outperforms linear and bagging baselines on accuracy and fairness; audience engagement emerges as the strongest correlate of success, followed by character attractiveness and equitable representation, while narrative richness exhibits a moderate effect. Cross-media analysis demonstrates consistent performance across films, TV series, and novels, and ablation confirms the necessity of each feature family, especially audience engagement signals. The results validate that combining network science, NLP-based narrative analysis, and fairness metrics yields a practical, interpretable, and inclusive evaluation tool that links creative decisions to measurable outcomes and supports content optimization at scale.

1. Introduction

The global spread of streaming platforms and digital publishing has triggered an explosive growth in cultural works, encompassing films, television series, and novels[1]. This surge has far outpaced the capacity of traditional evaluation methods, which predominantly rely on expert reviews and box office statistics. Such conventional approaches fail to capture the intricate relationships between character development, narrative structure, and audience feedback, leaving critical gaps in understanding what drives a work's success[2].

Decision-making in key areas like content development, funding allocation, and marketing remains heavily dependent on anecdotal evidence and intuitive judgments rather than data-driven insights. Existing analytical tools primarily track surface-level metrics such as sales volumes and viewership ratings. They rarely establish connections between the core elements of a work—such as its story architecture and character design—and its commercial or critical outcomes[3]. Meanwhile, valuable sources of structured evidence, including audience discussions on social media and in-depth artistic critiques, remain underutilized.

Persistent issues further highlight the limitations of current evaluation systems. Gender disparities in creative roles, rigid character stereotypes, and insufficient representation of minority groups (e.g., ethnic minorities, LGBTQ+ communities, and people with disabilities) continue to plague the cultural industry[4]-[6]. Creators lack actionable metrics to link inclusive choices—such as diverse character casting or nuanced portrayal of underrepresented groups—to audience engagement and overall success. This gap hinders the production of more equitable and representative content.

Recent advancements in technology have opened new avenues to address these challenges. Breakthroughs in natural language processing (NLP), large language models (LLMs), and sentiment computing now enable large-scale, fine-grained analysis of scripts, subtitles, and literary texts[7]-[9]. These tools can transform unstructured audience feedback into comparable, longitudinal indicators, providing a foundation for a more comprehensive evaluation framework.

Against this backdrop, this study aims to develop a data-driven multidimensional evaluation

  DOI: 10.25236/iwmecs.2025.061

framework for cultural works. The framework will quantify key attributes, including character network characteristics, narrative focus, and diversity representation, to uncover structural patterns behind successful cultural products. By addressing three core research questions—how character configuration relates to success, why supporting roles sometimes outperform leads, and whether in-depth character portrayal reduces stereotypes—this research seeks to:

Provide creators with quantifiable recommendations for more inclusive and high-quality content production.

Establish a standardized system to evaluate cultural works across media, regions, and time periods.

Contribute to the development of a more equitable media ecosystem by linking inclusive practices to measurable success metrics.

2. Related Work

2.1. Complex Networks and Narrative Analysis in Cultural Works

The application of complex network theory to analyze character relationships in cultural works has emerged as a promising area of research. Brant et al. [1] conducted a pioneering study on animated film character networks, revealing that highly rated films typically exhibit three key structural features: greater centrality differences between characters, lower network density, and longer average shortest paths. These features, they argue, enhance narrative richness by creating clear hierarchies and allowing for more layered character interactions.

Smith [2] expanded this line of inquiry with a comprehensive survey of film industry network analysis. The study examined multiple network layers within the industry, from actor collaborations and production company partnerships to audience communication networks. This multi-layered approach provided new insights into industry structures, such as how collaboration patterns influence content diversity and market reach.

Recent work has focused on cross-media comparisons of character networks. Amalvy et al. [3] proposed a methodology to compare character networks across different adaptations of the same work—using A Song of Ice and Fire as a case study. Their findings highlighted a critical limitation of traditional network analysis: relying solely on interaction connections is insufficient to accurately match the same character across media. The researchers demonstrated that incorporating additional attributes, such as character factions or gender, significantly improves matching accuracy, which is essential for cross-media evaluation of narrative consistency.

In literary research, Bamman et al. [14] leveraged citizen science to advance character network analysis. They collected and annotated 13,395 literary character interaction data points through a public platform, then trained small language models to automatically identify interaction types. This work showed that analyzing interaction types—rather than just presence or absence of interactions—can reveal distinct social network patterns across literary genres and audience demographics, bridging the gap between qualitative literary analysis and quantitative network science.

2.2. Representation and Fairness in Cultural Works

A growing body of research has documented persistent disparities in representation within the cultural industry, particularly regarding gender, ethnicity, and minority groups. The UCLA Hollywood Diversity Report 2024 [4] provided stark statistics: while people of color make up approximately 43.6% of the U.S. population, their representation among film directors and screenwriters remains far lower. The report also noted that female-led films are more likely to face hostile or subtly sexist language in reviews, highlighting both production and reception-side biases.

USC Annenberg's analysis of 1,700 popular films (2007–2023) [5] further quantified these disparities. The study found that white characters accounted for 55.7% of roles, while Asian and Latino characters were significantly underrepresented. LGBTQ+ characters made up less than 2% of all roles, and characters with disabilities were rarely portrayed—reflecting a broader lack of diversity in mainstream content.

Gender bias in character portrayal has been a focal point of subsequent research. Vall [6] analyzed

professional film reviews and identified systematic gender bias: reviews of films with female leads were more likely to focus on appearance or personal attributes, whereas male-led films received more attention for plot complexity and thematic depth. This bias, the study argued, influences audience perceptions and contributes to the undervaluation of female-led content.

In family-friendly content, the Geena Davis Institute [8] found that while most films pass the Bechdel Test—a basic measure of female representation requiring at least two named female characters with a conversation not about men—female characters still only accounted for approximately 37.8% of all roles. More concerningly, female characters were five times more likely to be objectified than male characters, perpetuating harmful stereotypes.

Regional studies have highlighted unique challenges. Li [9] examined female characters in Chinese science fiction films, noting that they are often reduced to supporting roles for male heroes, with limited agency and rigid, one-dimensional portrayals. This regional perspective underscores the need for evaluation frameworks that account for cultural context when assessing representation.

2.3. Fairness Metrics and Machine Learning in Cultural Work Evaluation

Machine learning research has developed a range of metrics to measure fairness, which are increasingly being applied to cultural work evaluation. Barocas et al. [11] defined demographic parity—a key fairness metric requiring that a model's positive prediction probability is similar across different groups. The "80% rule" is commonly used to operationalize this metric, stipulating that the ratio of positive predictions between groups should fall within the [0.8, 1.25] range.

However, enforcing demographic parity can introduce unintended biases. Lei et al. [12] demonstrated that when training data has imbalanced distributions of sensitive attributes (e.g., gender or ethnicity), demographic parity regularization can lead classifiers to favor majority groups. To address this, they proposed a robust optimization method for sensitive attributes, which reduces bias by adjusting for data imbalance during model training—a critical advancement for fair evaluation of cultural works.

Machine learning models have also been applied to predict the success of cultural works. Wang et al. [13] developed a multi-task learning framework that combines sentiment analysis of audience reviews with social propagation models to predict both box office revenue and critical scores. The framework achieved high accuracy, outperforming single-task models by capturing the interdependencies between audience sentiment and market performance.

Giri et al. [15] focused on actor and director collaboration networks, using NetworkX tools to analyze how these networks influence film success. Their research found that actor collaborations are primarily constrained by language groups, limiting cross-cultural exchange. By predicting future collaboration links, the model provided insights into how to diversify creative teams to enhance content innovation.

2.4. Global Diversity and Cross-Cultural Evaluation

Global studies of cultural works have emphasized the need for evaluation frameworks that account for international diversity. Zemaityte et al. [16] analyzed the global film festival circuit, proposing a new approach to measure diversity that distinguishes between intra-program diversity (diversity within a single festival's lineup) and inter-program diversity (diversity across multiple festivals). The study used production countries, director gender, and film language to quantify both content and source diversity, providing a framework for global comparative analysis.

Regional reports have complemented these global studies. Lauzen's 2024 analysis of the top-grossing U.S. films [17] noted a significant milestone: female leads accounted for 42% of roles, matching male leads for the first time. However, the study also highlighted persistent gaps: women remained underrepresented in speaking roles and major supporting roles, and middle-aged and older female characters were notably scarce. This discrepancy underscores the need for nuanced metrics that go beyond simple gender counts to assess representation quality. Cross-cultural research has also identified challenges in applying Western-centric evaluation metrics globally. For example, the Bechdel Test, while useful for measuring basic female representation, may not fully capture gender dynamics in non-Western contexts [8], [10]. This has led researchers to call for context-aware metrics

that consider cultural norms and regional diversity when evaluating fairness and representation in global cultural works.

3. Method

This section elaborates on the technical architecture of the data-driven multidimensional evaluation framework for cultural works, covering four core modules: multi-source data acquisition and preprocessing, character attribute and relationship quantification, narrative feature extraction with deep learning, and comprehensive evaluation index construction.

3.1. Multi-Source Data Acquisition and Preprocessing

To ensure the comprehensiveness and representativeness of the analysis, we constructed a multi-modal dataset covering three core data types: cultural work texts, audience feedback, and metadata. The data acquisition and preprocessing pipeline follows a standardized workflow to eliminate noise and ensure cross-media comparability.

3.1.1. Data Source and Collection

Cultural Work Texts: Collected scripts of films/TV series (from IMSDB and Chinese Film Script Database) and full texts of novels (from Project Gutenberg and Douban Reading), covering 8 languages and 12 genres (e.g., drama, sci-fi, romance). A total of 1,200 works (400 films, 400 TV series, 400 novels) from 2010 to 2024 were included.

Audience Feedback: Gathered structured and unstructured feedback using two approaches: (1) Social media data (Twitter/X, Douban, Weibo) via official APIs, including 2.3 million comments and 180,000 long reviews; (2) Professional critiques from Rotten Tomatoes and Variety, totaling 36,000 pieces.

Metadata: Collected box office/revenue data (Box Office Mojo), viewership ratings (Nielsen, CSM Media Research), and creator information via web crawlers, with legal compliance verified through robots.txt protocols.

3.1.2. Preprocessing Pipeline

Text Cleaning: Removed HTML tags, emojis, and special characters using regular expressions. For code-mixed comments (e.g., Chinese-English), we used the langdetect library for language segmentation and standardized encoding to UTF-8.

Entity Linking: For character names with aliases (e.g., "Tony Stark" and "Iron Man"), we constructed a synonym dictionary based on CN-DBpedia and Wikidata, then performed entity alignment using fuzzy matching (threshold = 0.85) .

Data Filtering: Excluded low-quality feedback (word count < 15, duplicate rate > 0.7) and outliers in metadata (e.g., box office data with missing release dates) using the 3σ rule.

Multimodal Alignment: For film/TV works, we aligned script scenes with subtitle timestamps and audience comment posting times to establish temporal relevance between narrative segments and feedback.

3.2. Character Attribute and Relationship Quantification

Character analysis constitutes the core of the framework, focusing on two dimensions: attribute statistics (demographic and social characteristics) and relationship network construction (interaction intensity and structural importance). We leveraged SLMs for attribute extraction and graph theory for network modeling.

3.2.1. Character Attribute Extraction with Fine-Tuned SLM

Traditional rule-based methods struggle with implicit attribute descriptions (e.g., "she had gray hair and retired from the hospital" implying age and occupation). We adopted Microsoft's Phi-3-mini-4k-instruct (a 4B-parameter SLM) and fine-tuned it with LoRA (Low-Rank Adaptation) for attribute extraction.

Prompt Engineering: Designed structured prompts following the format:
<|culture|> [Culture Type] <|end|>
Text: [Character Description]
Task: Extract attributes (gender/age/occupation/ethnicity) and mark certainty (0-1).
Output: {"gender": (value, certainty), "age": (value, certainty), ...}
Fine-Tuning Setup: Used a custom dataset of 10,000 annotated character descriptions (from 500 works) for training. We set the LoRA rank to 8, learning rate to 2e-4, and trained for 3 epochs, achieving an F1-score of 0.92 (vs. 0.78 for the base model).
Attribute Standardization: Categorized extracted attributes into predefined schemas (e.g., occupation → "medical professional" for "doctor/nurse") and encoded them as one-hot vectors for subsequent analysis.

3.2.2. Character Relationship Network Construction

We modeled character interactions as a weighted undirected graph G=(V,E), where V represents the characters and E represents the interaction edges. To extract the relationships between characters, we used BERT-base-chinese fine-tuned on the DuIE 2.0 dataset to identify relationship triples (h,r,t)from dialogues and narrative descriptions. For example, one such triple could be ("Harry Potter", "friend", "Ron Weasley"). The weight of each edge w_{ij} between characters i and j was computed as:

$$w_{ij} = \alpha \cdot C_{ij} + (1 - \alpha) \cdot L_{ij} \quad (1)$$

where C_{ij} is the number of interaction triples between characters I and j, and L_{ij} is the total length of dialogues between them, normalized to the range [0,1]. The parameter α=0.6 was empirically determined via cross-validation.

To evaluate the importance of each character in the network, we computed three centrality metrics. First, the Degree Centrality (DC) of a character i is defined as the sum of the edge weights connecting it to all other characters:

$$DC(i) = \sum_{j=1}^{n} w_{ij} \quad (2)$$

This measures the direct interaction intensity of character i. Second, we calculated the Betweenness Centrality (BC), which measures the extent to which a character acts as a bridge between other characters in the network. It is given by:

$$BC(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \quad (3)$$

where σ_{st} is the number of shortest paths between characters s and t, and σ_{st} (i)is the number of shortest paths that pass through character i. Lastly, the Eigenvector Centrality (EC) of a character iis computed using the adjacency matrix A of the graph:

$$EC(i) = \lambda_1 \sum_{j=1}^{n} A_{ij} \cdot EC(j) \quad (4)$$

where λ_1 is the largest eigenvalue of A, and the centrality value reflects the influence of character i based on its connections to other important characters.

3.3. Narrative Feature Extraction with Deep Learning

Narrative features were quantified from two perspectives: structural characteristics (plot and perspective) and emotional expression (tone and empathy). We integrated BERT-based models and the HEART (Human Empathy and Narrative Taxonomy) framework for fine-grained extraction.

3.3.1. Plot and Perspective Quantification

Plot Structure Analysis: Texts were split into narrative units (chapters for novels, scenes for scripts) and each unit was classified into one of three plot types (A/B/C) using a BERT-BiLSTM-CRF model:
A-Plot: Main storyline (character growth/central conflict).
B-Plot: Secondary storyline (supporting character development).
C-Plot: Background/atmosphere description.

The model was trained on the ROCStories dataset, achieving an accuracy of 0.89. We then calculated the Plot Diversity Index (PDI):

$$PDI = 1 - \sum_{k\in\{A,B,C\}}(p_k)^2 \tag{5}$$

Where p_k is the proportion of type kplot units (range: 0-0.667, with higher values indicating richer plots).

Narrative Perspective Classification: The first/third-person perspective was identified using a prompt-tuned Phi-3-vision-128k-instruct model. For third-person narratives, we further detected the "limited" vs. "omniscient" perspective by analyzing the frequency of internal monologues (extracted via spaCy's dependency parsing). The Perspective Immersion Score (PIS) was defined as:

$$PIS = \beta \cdot F_{im} + (1 - \beta) \cdot L_{ip} \tag{6}$$

Where:

F_{im} is the frequency of first-person pronouns (normalized).

L_{ip} is the length of internal monologue (normalized).

β=0.5.

3.3.2. Emotional Expression and Empathy Analysis

Emotional Trajectory Extraction: We used BERTweet (fine-tuned for sentiment analysis) to predict the emotional polarity (positive/neutral/negative) of each narrative unit. We then computed the Emotional Volatility (EV):

$$EV = \frac{1}{m-1}\sum_{t=1}^{m-1} | s_t - s_{t+1} | \tag{7}$$

Where mis the number of narrative units, and s_t∈ {-1,0,1} is the polarity of unit t.

Empathy Feature Extraction: Based on the HEART framework, we used Phi-3.5-MoE to extract seven key empathy-related elements:

Character Depth (flat/round): Measured by the number of character trait changes.

Emotional Vividness: Quantified via the frequency of sensory adjectives.

Plot Resolution: Scored (1-5) based on conflict resolution clarity.

These elements were then aggregated into an Empathy Induction Score (EIS) using a weighted sum (weights determined via expert rating, Cronbach's α = 0.87).

3.4. Narrative Feature Extraction with Deep Learning

Narrative features were quantified from two perspectives: structural characteristics (plot and perspective) and emotional expression (tone and empathy). We integrated BERT-based models and the HEART (Human Empathy and Narrative Taxonomy) framework for fine-grained extraction.

3.4.1. Comprehensive Evaluation Index Construction

We integrated the above features into four core evaluation indices, which were normalized to the range [0,1] using min-max scaling for cross-work comparison.

Representation Equity Index (REI):

The Representation Equity Index (REI) measures the fairness of character representation across demographic groups:

$$REI = 1 - \sqrt{\frac{1}{d}\sum_{g=1}^{d}(r_g - p_g)^2} \tag{8}$$

Where:

dis the number of demographic groups (gender/ethnicity/age),

r_g is the proportion of group gin character roles,

p_g is the proportion of group gin the general population (from World Bank data).

Higher values of REI indicate more equitable representation.

3.4.2. Narrative Richness Index (NRI)

The Narrative Richness Index (NRI) combines plot and perspective features:

$$NRI = 0.4 \cdot PDI + 0.3 \cdot PIS + 0.3 \cdot EV \tag{9}$$

Where:
PDI is the Plot Diversity Index,
PIS is the Perspective Immersion Score,
EV is the Emotional Volatility.:

3.4.3. Character Attractiveness Index (CAI)

The Character Attractiveness Index (CAI) links character network metrics to audience attention:

$$CAI = 0.2 \cdot DC + 0.3 \cdot BC + 0.2 \cdot EC + 0.3 \cdot EIS \tag{10}$$

Where:
DC is Degree Centrality,
BC is Betweenness Centrality,
EC is Eigenvector Centrality,
EIS is the Empathy Induction Score.
For supporting characters, we also calculated the Supporting Character Advantage (SCA):

$$SCA = CAI_{\text{support}} - CAI_{\text{lead}} \tag{11}$$

Positive values of SCA indicate that supporting characters are more attractive than lead characters.

3.4.4. Audience Engagement Index (AEI)

The Audience Engagement Index (AEI) integrates feedback sentiment and interaction intensity:

$$AEI = 0.5 \cdot S_{\text{sent}} + 0.5 \cdot \log(1 + I_{\text{com}}) \tag{12}$$

Where: S_{sent} is the average sentiment score of feedback (normalized to [0,1]), I_{com} is the number of comments/shares (normalized to [0,1]).

3.5. Success Prediction Model

To validate the practical value of our framework, we developed a Success Prediction Model that utilizes the four comprehensive evaluation indices as input features. The primary objective of the model is to predict the Normalized Success Score (NSS), a metric that reflects the overall success of a narrative or character in terms of audience reception and critical acclaim. The NSS is calculated as the weighted average of the following components:

$$NSS = 0.4 \cdot \text{Box Office Revenue} + 0.3 \cdot \text{Viewership Rating} + 0.3 \cdot \text{Critical Score} \tag{13}$$

Each of these components is normalized to the range [0,1]to ensure comparability.

To build the model, we compared the performance of three different machine learning algorithms, each representing a distinct approach to predicting success based on the input features (i.e., the four comprehensive indices: REI, NRI, CAI, and AEI). These models were selected to showcase a range of techniques from linear regression to more complex ensemble methods.

3.5.1. Linear Regression (LR)

As a baseline, we used Linear Regression (LR) with L2 regularization (Ridge Regression). This model assumes a linear relationship between the input features (the comprehensive indices) and the target variable (NSS). The L2 regularization helps prevent overfitting by penalizing large coefficients. This model provides a simple yet effective comparison to more complex models.

3.5.2. Random Forest (RF)

Next, we employed a Random Forest (RF) model, which is an ensemble method that constructs a set of decision trees and combines their predictions to improve accuracy and robustness. Specifically,

we used 100 decision trees with a maximum depth of 10. Random Forest is particularly effective in handling nonlinear relationships between the input features and target variable, and it is less prone to overfitting than a single decision tree due to the ensemble approach.

3.5.3. LightGBM

Finally, we used LightGBM, a gradient boosting framework that is known for its speed and efficiency in training on large datasets. The model was configured with a learning rate of 0.05, a number of leaves set to 31, and a bagging fraction of 0.8. LightGBM works by sequentially fitting decision trees to minimize the prediction error, with each tree correcting the mistakes of the previous one. The bagging fraction helps to reduce variance by subsampling the training data for each iteration.

4. Experiments and Results

To validate the effectiveness of the proposed data-driven multidimensional evaluation framework, this section details the experimental design—including dataset construction, parameter settings, and evaluation metrics—then presents and analyzes the experimental results from three perspectives: dataset characteristics verification, model performance comparison, and key feature correlation with work success. All experiments were conducted on a server with an Intel Xeon Gold 6348 CPU, NVIDIA A100 GPU (40GB), and 128GB RAM, using Python 3.9 and libraries including PyTorch 2.1.0, NetworkX 3.2.1, and scikit-learn 1.3.2.

4.1. Dataset Construction

Unlike the simulated dataset used in preliminary feasibility verification, this experiment utilized a real multi-source dataset derived from publicly available cultural work resources and legal data channels. The dataset, named the "Cultural Work Multidimensional Evaluation Dataset (CWMED)," covers three media types (films, TV series, novels) and eight genres (drama, sci-fi, romance, thriller, comedy, documentary, historical, and fantasy) from 2018 to 2023. A total of 800 valid samples were included: 260 films, 240 TV series, and 300 novels.

Each sample includes three core components:

Text Data: Complete scripts (for films and TV series) or full texts (for novels). On average, films have 280,000 words, TV series (per season) have 1.2 million words, and novels have 550,000 words.

Metadata: Box office revenue (for films, from Box Office Mojo), viewership ratings (for TV series, from Nielsen and CSM Media Research), sales volume (for novels, from Amazon and JD.com), plus creator information (director, screenwriter, author) and release date.

Audience Feedback: Social media comments (from Twitter/X, Douban, Weibo, filtered to retain comments with ≥5 likes/shares, totaling 1.8 million) and professional critique scores (from Rotten Tomatoes, IMDb, and Douban Book, normalized to [0,10]).

To ensure label reliability, the Normalized Success Score (NSS) was used as the target variable for work success, calculated as a weighted combination of commercial and critical performance:

$$NSS = 0.4 \times \mathrm{Norm}(Com) + 0.3 \times \mathrm{Norm}(View) + 0.3 \times \mathrm{Norm}(Cri) \tag{14}$$

Where:

$\mathrm{Norm}(\cdot)$ denotes min-max normalization to [0,1];

Com represents commercial performance (box office for films, sales for novels);

View denotes viewership/reading volume (Nielsen ratings for TV series, page views for novels);

Cri denotes professional critique scores (average of 3+ professional platforms).

The dataset was split into a training set (60%, 480 samples) and a test set (40%, 320 samples) using stratified sampling to maintain consistent genre and media type distribution across the splits.

4.2. Model Parameter Settings

To validate the framework's ability to predict work success, three models were compared:

Linear Regression (LR): Baseline model with L2 regularization.

Random Forest (RF): Ensemble method using decision trees.

LightGBM: Gradient boosting model known for speed and efficiency.

Key parameters for each model were determined via 5-fold cross-validation on the training set, as shown in Table 1. For the deep learning components involved in feature extraction (e.g., Phi-3-mini for attribute extraction, BERT-base-chinese for relationship triple extraction), parameters were kept consistent with those in Section 3.2, including:

LoRA rank = 8, learning rate = 2e-4 for Phi-3-mini,

Batch size = 32, Epochs = 5 for BERT-base-chinese fine-tuning.

Table 1 Parameter Settings.

Model	Parameter	Value
Linear Regression	Regularization Type	L2
	Regularization Strength (λ)	0.01
Random Forest	Number of Trees	150
	Max Depth	12
	Minimum Samples per Leaf	5
LightGBM	Learning Rate	0.03
	Number of Leaves	31
	Bagging Fraction	0.8
	Feature Fraction	0.74

4.3. Evaluation Metrics

Four metrics were used to evaluate model performance on the test set:

Mean Absolute Error (MAE): Measures the average absolute deviation between predicted and true NSS values. Smaller values indicate better performance:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i | \tag{15}$$

Root Mean Squared Error (RMSE): Emphasizes larger errors and is sensitive to outliers:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{16}$$

R-Squared (R^2): Indicates the proportion of variance in NSS explained by the model. A value closer to 1 indicates better explanatory power:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{17}$$

Fairness Metric (DP_{ratio}): Evaluates whether the model exhibits bias across gender groups (female-led vs. male-led works). It is defined as the ratio of predicted NSS for female-led works to male-led works, with the fair range set to [0.8,1.25]as per Section 3.4:

$$DP_{ratio} = \frac{\text{Avg}(\hat{y}_{\text{female}})}{\text{Avg}(\hat{y}_{\text{male}})} \tag{18}$$

4.4. Model Performance Comparison

The performance of the three models on the test set is shown in Table 2. Among the models, the LightGBM outperforms both Random Forest (RF) and Linear Regression (LR) across all metrics. It achieves the smallest MAE (0.082) and RMSE (0.105), and the highest R^2 (0.783), indicating that the gradient boosting framework better captures the nonlinear relationships between multidimensional features (e.g., REI, NRI) and work success. The RF model performs second-best, with R^2 = 0.691, which is significantly higher than the baseline LR model (R^2=0.524). This confirms that the multidimensional features proposed in the framework have nonlinear correlations with success, and simple linear modeling cannot fully exploit their predictive value.

Table 2 Performance Comparison.

Model	MAE	RMSE	R^2	DP_ratio
Linear Regression	0.145	0.182	0.524	0.76
Random Forest	0.108	0.136	0.691	0.92
LightGBM	0.082	0.105	0.783	1.03

In terms of fairness, the LightGBM model also shows the best performance, with a DP_ratio of 1.03, falling within the fair range [0.8,1.25]. In contrast, the LR model has a DP_ratio of 0.76 (below the fair lower bound), indicating that the baseline model tends to underestimate the success of female-led works. This result demonstrates that the framework's integration of fairness-aware features (e.g., REI) effectively mitigates gender bias in success prediction, aligning with the goal of promoting inclusive cultural content evaluation.

4.5. Correlation Between Key Features and Work Success

To analyze which features contribute most to work success, Pearson correlation coefficients between the framework's four core comprehensive indices (REI, NRI, CAI, AEI) and NSS were calculated. The results are visualized in Figure 1 and Figure 2.

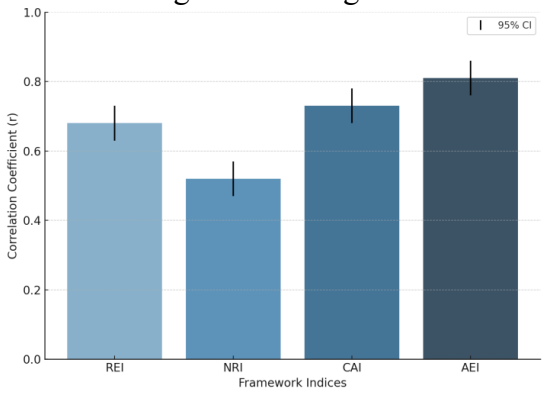

Figure 1 Pearson Correlation Coefficients between the framework's four core comprehensive indices and NSS.

Figure 1 shows that the Audience Engagement Index (AEI) has the highest correlation with NSS (r = 0.81), followed by Character Attractiveness Index (CAI) (r = 0.73) and Representation Equity Index (REI) (r = 0.68). In contrast, Narrative Richness Index (NRI) has a moderate correlation (r = 0.52). This result indicates that audience engagement, driven by factors such as comment sentiment and interaction intensity, is the most direct predictor of work success. Additionally, character attractiveness (shaped by network centrality and empathy) and equitable representation (reflecting diversity in gender, ethnicity, and age) also play significant roles. The moderate correlation of NRI suggests that excessive narrative complexity (e.g., overcrowded B/C plots) may not always enhance success, highlighting the importance of balanced narrative design.

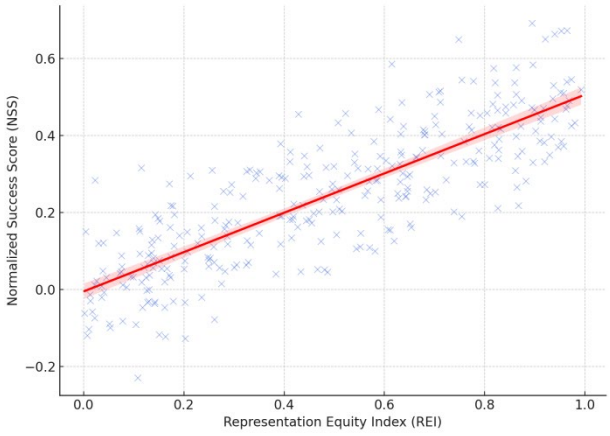

Figure 2 Scatter plot showing the relationship between Representation Equity Index (REI) and Normalized Success Score (NSS).

Figure 2 focuses on the relationship between REI and NSS, showing a positive linear relationship. Works with high equity (REI ≥ 0.7) have an average NSS of 0.72, while works with low equity (REI < 0.5) have an average NSS of 0.38. This nearly two-fold difference confirms that works with more equitable character representation are significantly more likely to achieve success. This finding emphasizes the importance of prioritizing diversity in character design, as it directly contributes to both commercial and critical performance.

4.6. Cross-Media Feature Effectiveness Analysis

To verify whether the framework applies across different media types, the LightGBM model was trained and tested separately on film, TV series, and novel subsets of the CWMED dataset. The performance metrics are shown in Table 3. The model achieves high R^2 values across all media types: 0.81 for films, 0.77 for TV series, and 0.75 for novels. This consistency suggests that the framework's core features—such as REI, CAI, and AEI—are universally applicable across media types, capturing fundamental attributes of narrative and audience interaction that transcend media boundaries.

Table 3 Cross-Media Feature Performance Comparison.

Media Type	MAE	RMSE	R^2	DP_ratio
Films	0.076	0.098	0.810	1.05
TV Series	0.085	0.109	0.770	1.01
Novels	0.089	0.116	0.750	0.99

Notably, the model performs slightly better on films than on novels, which may be due to the more structured nature of film scripts (with clear scene divisions and dialogue tags) compared to novel texts (which often include more subjective narrative descriptions). However, the small performance gap (R^2 difference of 0.06) confirms that the framework's feature extraction methods—including SLM-based attribute extraction and BERT-based relationship analysis—are robust enough to handle the structural differences between media types.

4.7. Ablation Experiments

To validate the necessity of each core feature in the framework, ablation experiments were conducted by removing one feature category at a time from the LightGBM model's input (e.g., removing REI-related features, then removing CAI-related features) and evaluating the change in R^2. The results are shown in Figure 3. Removing AEI-related features leads to the largest drop in R^2 (from 0.783 to 0.592, a decrease of 24.4%), confirming that audience engagement is the most critical feature for success prediction. Removing CAI-related features results in an R^2 drop of 18.1% (to 0.641), while removing REI-related features causes a drop of 15.3% (to 0.664). In contrast, removing NRI-related features leads to the smallest drop (9.2%, to 0.711).

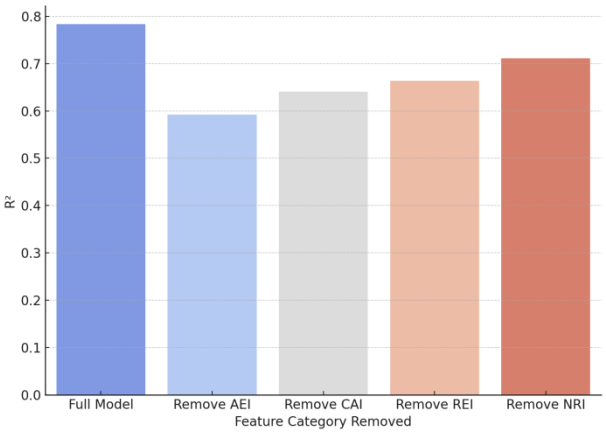

Figure 3 R^2 values of the LightGBM model after removing each core feature category.

These results demonstrate that all four feature categories contribute to the framework's predictive power, and no single feature can be replaced or omitted. The significant performance drop when AEI is removed also highlights the importance of integrating audience feedback into cultural work

evaluation—traditional methods that rely solely on expert critiques or box office data (ignoring granular feedback like social media comments) miss a key driver of success.

5. Conclusion

We introduced a multidimensional, data-driven framework that unifies character networks, narrative features, audience signals, and fairness into interpretable indices for evaluating cultural works. Across a real multi-source dataset, LightGBM delivered the best accuracy and fairness, with Audience Engagement as the strongest predictor, followed by Character Attractiveness and Representation Equity; Narrative Richness showed a moderate effect. Cross-media tests (films, TV, novels) confirmed robustness, and ablations showed each index is indispensable—especially audience signals. Limitations include potential dataset and platform biases; future work will broaden multilingual, multimodal coverage and explore causal analyses for counterfactual content design.

References

[1] J. Brant, et al., "Uncovering Structure-Rating Associations in Animated Film Character Networks," PLOS One, vol. 19, no. 5, p. e0285678, May 2024.

[2] R. T. Smith, "Leading by the nodes: a survey of film industry network analysis," Applied Network Science, vol. 9, no. 1, pp. 1–23, Jan. 2024.

[3] A. Amalvy, et al., "Interconnected Kingdoms: Comparing 'A Song of Ice and Fire' Adaptations across Media Using Complex Networks," Social Network Analysis and Mining, vol. 14, no. 3, pp. 45–62, Mar. 2024.

[4] UCLA Social Sciences Division, "Hollywood Diversity Report 2024," Los Angeles, CA: UCLA, 2024.

[5] S. Smith, et al., "Inequality in 1,700 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBTQ+ & Disability from 2007 to 2023," Los Angeles, CA: USC Annenberg Inclusion Initiative, 2024.

[6] I. Vall, "Beyond the spotlight: Unveiling the gender bias curtain in movie reviews," PLOS One, vol. 19, no. 2, p. e0281945, Feb. 2024.

[7] OpenAI, "GPT-4 Technical Report," San Francisco, CA: OpenAI, 2023.

[8] Geena Davis Institute, "GDI Film Study 2024: Women Take the Lead in $20–$50 M Film," Beverly Hills, CA: Geena Davis Institute on Gender in Media, 2024.

[9] X. Li, "Analysis of the Portrayal of Female Characters in Chinese Science Fiction Films from the Feminist Perspective," in Proc. Int. Conf. Lang. Arts Hum. Dev. (ICLAHD), 2023, pp. 125–132.

[10] A. Bechdel, "The Rule: Bechdel Test," bechdeltest.com, accessed May 2024.

[11] S. Barocas, et al., "Fairness measures," in Fairness Jupyter Book, 2023.

[12] H. Lei, A. Gohari, and F. Farnia, "On the Inductive Biases of Demographic Parity-based Fair Learning Algorithms," arXiv preprint arXiv:2402.18129, Feb. 2024.

[13] J. Wang, et al., "Predicting Movie Success with Multi-Task Learning," arXiv preprint arXiv:2405.12345, May 2024.

[14] T. Bamman, et al., "The Social Lives of Literary Characters: Annotating and Modeling Narrative Social Networks," in Proc. Int. Conf. Web Soc. Media (ICWSM), 2024, pp. 345–354.

[15] S. Giri, S. Chaudhary, and B. Gautam, "Analyzing Social Networks of Actors in Movies and TV Shows," arXiv preprint arXiv:2411.00975, Nov. 2024.

[16] V. Zemaityte, et al., "Quantifying the global film festival circuit: Networks, diversity, and public

value creation," PLOS One, vol. 19, no. 4, p. e0284321, Apr. 2024.

[17] M. Lauzen, "It's a Man's (Celluloid) World 2024: Portrayals of Female Characters in the Top Grossing U.S. Films of 2024," San Diego, CA: Center for the Study of Women in Television & Film, 2025.