ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

MACHINE LEARNING AND AUDIENCE INTELLIGENCE FOR MOVIE BUSINESS FORECASTING: A REVIEW

Aishwarya M Patil, Arpit Solanki
Dr. A.P.J. Abdul Kalam University, Indore, M.P., India

KEYWORDS: Movie Business Forecasting; Box-Office Prediction; Machine Learning; Sentiment Analysis; Social Media Analytics; Audience Intelligence; Ensemble Learning; Temporal Validation.

ABSTRACT

Forecasting the commercial performance of a film is a complex task because box-office outcomes are shaped by interacting production, market and audience-related factors. Traditional studies examined variables such as genre, budget, star power and release timing, whereas the growth of digital platforms has made social-media activity, online ratings and review sentiment available as additional sources of predictive information. This review surveys the development of movie business forecasting, focusing on statistical foundations, machine-learning methods, audience intelligence and ensemble-based prediction. The literature suggests that online signals can complement structured film attributes, but their contribution depends on data quality, platform characteristics, review volume and the time at which the information becomes available. Differences in target definitions and evaluation designs also make the results of separate studies difficult to compare. Three research gaps are identified from the literature: sentiment is usually summarised without recording how many reviews support it or how strongly reviewers disagree; prediction cutoffs and validation procedures differ across studies; and the contribution of individual model components is seldom isolated. The future scope outlines several directions, including reliability-aware adaptive forecasting, rigorous chronological evaluation and multi-channel forecasting across theatrical and streaming distribution.

INTRODUCTION

The motion-picture industry requires decisions to be made before the final response of audiences is known. Producers and studios estimate the commercial potential of projects while planning investment, production scale, casting and promotion. Distributors and exhibitors must anticipate demand when deciding release dates, screen allocation and marketing expenditure. These decisions involve considerable uncertainty because a film's revenue cannot be determined by one factor alone. Genre, budget, cast, director, runtime and release timing may influence performance, but audience expectations, competition, reviews and word-of-mouth can change demand during the release cycle. Early empirical work examined the relationship between film attributes and commercial outcomes [1], [2].

Digital communication has expanded the information available for forecasting. Social networking services, search activity, online ratings and written reviews provide observable traces of audience attention and opinion. These signals can be collected before or during a film's theatrical run and may complement conventional metadata. Asur and Huberman [3] demonstrated the predictive potential of social-media discussion, while Kim et al. [4] investigated machine-learning approaches using social networking service data. Later studies compared the predictive role of different platforms and examined review sentiment as a source of information about audience response [5], [6].

However, the presence of more data does not automatically produce a more dependable forecast. A large volume of online discussion may reflect publicity or controversy rather than positive demand. Sentiment scores can be unstable when only a few reviews are available, and disagreement among reviewers may be hidden by an average score. Moreover, a model intended for pre-release use must not rely on information generated after the forecast cutoff. If final ratings, actual opening revenue or post-release popularity are included, the evaluation may no longer represent a genuine pre-release prediction.

This review examines the development of movie business forecasting from statistical models to machine-learning and audience-intelligence approaches. It compares the types of information used, the contributions of major methods and the limitations that affect their practical use. The discussion also considers reliability-aware

forecasting as a future research direction. The review is thematic and selective; it is not presented as an exhaustive PRISMA systematic review.

RELATED REVIEW

2.1 Traditional Film Attributes and Economic Foundations

Early research on movie success primarily relied on structured information about films and the market. Litman [1] examined factors associated with theatrical success, while Ravid [2] studied the role of information, stars and blockbusters in the film industry. De Vany and Walls [7] emphasised the highly uncertain and uneven distribution of box-office revenue, showing that a small number of films can account for a substantial share of total receipts. Basuroy et al. [8] further examined the joint influence of critical reviews, star power and budgets. Collectively, these studies established that film performance is multi-factorial and that the presence of well-known actors or large budgets does not eliminate commercial uncertainty.

Structured attributes remain useful because they are generally available before release and can provide a stable baseline for prediction. Variables such as genre, production budget, runtime, cast and crew information, release month and historical popularity can describe important aspects of a film and its market position. Nevertheless, these variables may not fully represent the changing expectations of potential viewers. This limitation encouraged researchers to investigate digital traces of audience attention and opinion.

2.2 Social Media and Electronic Word-of-Mouth

Social media provides rapidly updated information about public attention. The number of posts, mentions, shares and discussions may indicate awareness and interest in an upcoming film. Asur and Huberman [3] showed that social-media activity can be useful for predicting movie-related outcomes. Kim et al. [4] subsequently used social network service data with machine-learning algorithms, illustrating how online variables can be integrated with predictive models. Such work shifted the focus from static film characteristics towards signals that change as a release approaches.

Online word-of-mouth is not a uniform measure. Different platforms attract different user groups and encourage different styles of interaction. Bogaert et al. [5] compared the predictive ability and mechanisms of social-media platforms, highlighting the importance of considering platform-specific behaviour rather than treating all online activity as equivalent. Temporal patterns are also relevant: a sudden increase in discussion close to release may have a different meaning from sustained interest over several weeks. Najafi [9] examined temporal tweet patterns for motion-picture box-office prediction, supporting the value of considering how online attention evolves over time. Thus, both the volume and timing of audience activity may matter.

Despite its potential, social-media information has limitations. Online discussion may be affected by paid promotion, bots, repeated posts, fan communities or controversies. Mention counts alone do not reveal whether audiences are expressing approval, dissatisfaction or simple curiosity. Platform access and user demographics may also vary across regions. Consequently, social-media features should be interpreted as indicators of online attention, not as direct measurements of ticket purchases.

2.3 Review Sentiment and Audience Response

Review-based forecasting attempts to extract the evaluative content of audience feedback. Sentiment analysis assigns a polarity or score to text, commonly representing positive, neutral or negative opinion. Such scores can be aggregated at film level and combined with structured attributes. Hur et al. [6] investigated the use of sentiments from movie reviews for box-office forecasting, while Chen and Dai [10] explored mining movie box-office and review topics from social-network big data. These approaches demonstrate that reviews may contain both overall emotional orientation and more specific information about audience concerns.

An average sentiment score, however, does not describe the strength of the evidence behind it. Two films may have the same average sentiment even if one has a large number of consistent reviews and the other has only a few highly conflicting opinions. Review count and sentiment dispersion can therefore provide useful contextual information. A reliability-aware representation can treat the amount of evidence and the degree of disagreement as separate features, rather than assuming that every sentiment estimate is equally dependable.

Sentiment-based methods also face language and context challenges. Sarcasm, slang, mixed-language comments, spoilers and domain-specific expressions may affect automated classification. Reviews posted after a film's release

ISSN 2349-4506
Impact Factor: 3.799



Global Journal of Engineering Science and Research Management

may not be suitable for a pre-release forecast unless the intended prediction date explicitly allows them. These issues make the definition of a prediction cutoff and the recording of review timestamps important parts of the forecasting design.

2.4 Machine Learning, Ensembles and Recent Developments

Machine-learning methods are widely used because movie data may contain nonlinear relationships and interactions that are difficult to represent with simple linear models. Regression models provide interpretable baselines, while Random Forest can capture nonlinear patterns through collections of decision trees. Support Vector Regression (SVR) can model nonlinear relationships through kernel functions. Neural networks offer additional flexibility when datasets are sufficiently large and features are appropriately represented. Mbunge et al. [11] reviewed trends in box-office prediction and the computational models used in the field, illustrating the range of approaches investigated by researchers.

Ensemble learning combines predictions from multiple models in an attempt to reduce the weaknesses of individual learners. A simple average gives each model equal influence, whereas validation-based weighting assigns greater importance to models that perform better on validation data. Adaptive combinations can go further by allowing the weight to vary with the characteristics or quality of the evidence available for a particular film. Qiu and Zheng [12] examined sentiment-informed forecast combinations and reported that regularisation-based combinations can improve box-office projections in their study setting. Zhang et al. [13] combined neural-network techniques with emotional mining, demonstrating another route for incorporating comment-derived information. Recent methodological discussion has also emphasised the need to interpret reported performance carefully. Leukel et al. [14] conducted a systematic review of machine-learning studies on movie box-office revenue prediction and found that many abstracts omit performance metrics and that strong or leading language is sometimes used without adequate quantitative support. This reinforces the importance of presenting evaluation metrics, baselines, data partitions and limitations clearly. A model's performance should be interpreted in relation to the market, prediction horizon and target used in the experiment.

2.5 Evaluation, Temporal Validity and Forecasting Reliability

Forecasting differs from explaining historical associations. In a real forecasting setting, a model must use only information available at the time the prediction is issued. Randomly splitting films into training and test sets can allow information from later releases to influence a model evaluated on earlier releases, a form of data leakage [15]. Chronological partitioning is therefore more appropriate when the intended use is to forecast future films [16]. The prediction cutoff should also be specified for each feature so that post-release information is not inadvertently included.

Evaluation metrics provide different views of prediction quality. Mean Absolute Error (MAE) describes the average absolute prediction error, while Root Mean Squared Error (RMSE) places greater emphasis on large errors. The coefficient of determination (R^2) indicates how much variation is explained relative to a mean-based reference, but it should not be interpreted in isolation. Revenue distributions are often highly skewed, so the target transformation and conversion back to the original monetary scale should be documented. Differences in target definition, currency, market, time window and evaluation split can make headline results across studies non-comparable.

Reliability should therefore be considered at more than one level. First, the input evidence may vary in quality, especially when review volume is low or opinions are highly dispersed. Second, model performance may vary across release periods, genres and markets. Third, a single point estimate does not communicate the uncertainty that decision-makers face. Repeated chronological evaluation, component-level ablation, comparison with multiple baselines and uncertainty intervals can help determine whether an apparent improvement is stable and practically meaningful.

COMPARATIVE ANALYSIS

The selected studies represent complementary approaches to movie business forecasting. Their results should not be compared as if they were obtained under an identical experimental setting, because the studies differ in data sources, target variables, markets and prediction horizons. The following comparison therefore focuses on their methodological contribution and the limitations that remain relevant.

ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

Table 3.1: Comparative Analysis of Existing Studies

Study	Primary information / approach	Main contribution	Important consideration
Litman [1]; Ravid [2]	Film attributes and economic factors	Established early determinants of movie success	Static attributes cannot fully capture changing audience response
De Vany & Walls [7]	Revenue distribution and uncertainty	Highlighted high uncertainty and uneven revenue outcomes	Predictive risk remains even when star power is considered
Asur & Huberman [3]; Kim et al. [4]	Social-media / SNS activity with predictive models	Introduced online attention as a forecasting signal	Volume may reflect promotion or curiosity, not positive demand
Bogaert et al. [5]; Najafi [9]	Cross-platform and temporal social-media patterns	Emphasised platform differences and time-dependent signals	Platform behaviour and user composition may limit transferability
Hur et al. [6]; Chen & Dai [10]	Review sentiment and topic mining	Extracted evaluative information from audience text	Text quality, language and review timing affect reliability
Mbunge et al. [11]	Review of ML prediction approaches	Summarised model families and research trends	Studies use different datasets and evaluation protocols
Qiu & Zheng [12]; Zhang et al. [13]	Sentiment combinations and neural emotional mining	Explored ways to combine audience emotion with forecasting	Benefits require validation across datasets and markets
Leukel et al. [14]	Systematic review of reporting practices	Highlighted the importance of accurate performance reporting	Consistent reporting and reproducibility remain necessary

The comparison indicates a gradual expansion of the forecasting information base. Traditional film attributes offer structured and often pre-release information, while social-media variables represent public attention and review sentiment captures the evaluative direction of audience response. Machine-learning models can combine these sources, and ensemble methods can reduce dependence on a single learner. Nevertheless, a reliable model must distinguish attention from approval, account for the strength of the evidence and preserve the time order of information. The main opportunity is therefore not simply to add more features or more complex algorithms, but to make the forecasting process more reliable and more closely aligned with the actual decision time.

RESEARCH GAP IDENTIFICATION AND PROBLEM STATEMENT

Research gap. Existing studies have improved movie business forecasting using film metadata [1], [2], [7], [8], social-media activity [3]–[5], [9], review sentiment [6], [10] and machine-learning models [4], [6], [11]–[13]. However, among the studies reviewed here, sentiment is mostly used as an aggregated score without considering review volume or reviewer disagreement [6], [10], [12], [13], and prediction cutoffs and validation procedures differ across studies, with a risk of data leakage [11], [14]–[16]. These gaps make reported results difficult to compare and may overstate real pre-release forecasting performance.

Problem identification. The problem identified from the literature is how to forecast movie business performance using structured film information and audience-generated signals, while accounting for the reliability of those signals and following a leakage-resistant chronological evaluation.

CONCLUSION AND FUTURE SCOPE

This review examined studies on movie business forecasting based on film metadata, social-media activity, review sentiment and machine-learning models. The literature shows that these sources can improve forecasting, but audience signals are not equally reliable, and inconsistent prediction cutoffs and validation procedures limit comparability. Further progress therefore depends as much on the reliability of audience evidence and on evaluation design as on adding new features or more complex algorithms.

ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

Future scope. Future work can address these gaps in several ways. One possible direction is a reliability-aware forecasting approach, in which the influence of sentiment is adjusted according to review volume and reviewer agreement. Other directions include evaluation with clear chronological cutoffs, ablation and multiple baselines, extension to multi-channel forecasting across theatrical and streaming distribution, and improved audience-signal modelling using transformer-based sentiment models.

REFERENCES

1. B. R. Litman, "Predicting success of theatrical movies: An empirical study," The Journal of Popular Culture, vol. 16, no. 4, pp. 159–175, 1983, doi: 10.1111/j.0022-3840.1983.1604_159.x.
2. S. A. Ravid, "Information, blockbusters, and stars: A study of the film industry," Journal of Business, vol. 72, no. 4, pp. 463–492, 1999, doi: 10.1086/209624.
3. S. Asur and B. A. Huberman, "Predicting the future with social media," in Proc. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp. 492–499, doi: 10.1109/WI-IAT.2010.63.
4. T. Kim, J. Hong, and P. Kang, "Box office forecasting using machine learning algorithms based on SNS data," International Journal of Forecasting, vol. 31, no. 2, pp. 364–390, 2015, doi: 10.1016/j.ijforecast.2014.05.006.
5. M. Bogaert, M. Ballings, D. Van den Poel, and A. Oztekin, "Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms," Decision Support Systems, vol. 147, Art. no. 113517, 2021, doi: 10.1016/j.dss.2021.113517.
6. M. Hur, P. Kang, and S. Cho, "Box-office forecasting based on sentiments of movie reviews and Independent subspace method," Information Sciences, vol. 372, pp. 608–624, 2016, doi: 10.1016/j.ins.2016.08.027.
7. A. De Vany and W. D. Walls, "Uncertainty in the movie industry: Does star power reduce the terror of the box office?," Journal of Cultural Economics, vol. 23, no. 4, pp. 285–318, 1999, doi: 10.1023/A:1007608125988.
8. S. Basuroy, S. Chatterjee, and S. A. Ravid, "How critical are critical reviews? The box office effects of film critics, star power, and budgets," Journal of Marketing, vol. 67, no. 4, pp. 103–117, 2003, doi: 10.1509/jmkg.67.4.103.18692.
9. H. Najafi, "Predicting motion picture box office performance using temporal tweet patterns," International Journal of Intelligent Computing and Cybernetics, vol. 11, no. 1, pp. 64–80, 2018, doi: 10.1108/IJICC-04-2017-0033.
10. Y. Chen and Z. Dai, "Mining of movie box office and movie review topics using social network big data," Frontiers in Psychology, vol. 13, Art. no. 903380, 2022, doi: 10.3389/fpsyg.2022.903380.
11. E. Mbunge, S. G. Fashoto, and H. Bimha, "Prediction of box-office success: A review of trends and machine learning computational models," International Journal of Business Intelligence and Data Mining, vol. 20, no. 2, pp. 192–207, 2022, doi: 10.1504/IJBIDM.2022.120825.
12. Y. Qiu and Y. Zheng, "Improving box office projections through sentiment analysis: Insights from regularization-based forecast combinations," Economic Modelling, vol. 125, Art. no. 106349, 2023, doi: 10.1016/j.econmod.2023.106349.
13. Z. Zhang, Y. Meng, and D. Xiao, "Prediction techniques of movie box office using neural networks and emotional mining," Scientific Reports, vol. 14, Art. no. 21209, 2024, doi: 10.1038/s41598-024-72340-z.
14. J. Leukel, Z. Liu, and V. Sugumaran, "Potential overinterpretation of results in the abstracts of machine learning studies for movie box office revenue prediction: A systematic review," Journal of Revenue and Pricing Management, vol. 25, pp. 368–377, 2026, doi: 10.1057/s41272-026-00574-9.
15. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, Art. no. 15, 2012, doi: 10.1145/2382577.2382579.
16. C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," Information Sciences, vol. 191, pp. 192–213, 2012, doi: 10.1016/j.ins.2011.12.028.