A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics

Luc Marcel Pellinger, 58611

Work project carried out under the supervision of:

Sreyaa Guha

17/12/2024

Abstract

This thesis investigates factors influencing movie success using deep learning techniques. It compares weak supervision to supervised deep learning approaches when predicting movie success after the box office release. A key focus of this study is to establish whether the multiple instance learning (MIL) approach can perform more accurate predictions using multivariate time series data from the post-release stage of movies. Additionally, it is assessed whether a MIL-based architecture delivers more interpretable results compared to supervised architectures. By integrating diverse data sources and features, this study provides a comprehensive perspective on how different deep learning techniques can be used to measure audience engagement as a time series to classify box office success. The findings advance academic understanding of the applicability of MIL in the movie domain and offer insights for industry stakeholders aiming to enhance deep learning architecture selection.

Keywords: Multivariate Time Series Classification, Movie Success Prediction, Model Comparison, Multiple Instance Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

The film industry uniquely blends cultural importance with economic impact (Dellarocas, Zhang, & Awad, 2007). For instance, Steven Spielberg's 2004 film *The Terminal* exemplifies how audience reception can significantly affect box office performance. Despite a substantial $110 million budget and a star-studded cast featuring Tom Hanks, the movie earned only $77 million, falling well short of expectations. Analysts attributed this underperformance largely to negative word-of-mouth from consumers, underscoring the critical role of audience-driven perceptions in determining financial outcomes.

Online reviews and ratings have further amplified the importance of consumer perceptions in the movie industry (Pang, Liu, & Golder, 2022). Online reviews are pivotal in shaping how films are evaluated and received by the mass audience. Studies show that 88% of Americans consider online reviews to be trustworthy and beneficial, associating them with improved reliability (Watson & Wu, 2022). For movie enthusiasts, these critiques serve as an essential decision-making tool, guiding their viewing choices and shaping their confidence in the movie they select (Loupos, Peng, Li, & Hao, 2023). The influence of reviews extends beyond audiences; exhibitors are also affected, as evidenced by a positive correlation between review ratings and the number of screens allocated to a film (Wang & Guo, 2017). This highlights the broader impact of consumer feedback on the industry's supply chain dynamics.

Economically, the sector is a major driver of employment and revenue, with over 400,000 individuals working in the U.S. motion picture and video game industries as of 2023 (Bureau of Labor, 2023). This highlights the movie industry's role as both an economic driver and a cultural influence in modern society.

The movie industry provides a rich data environment encompassing diverse data structures. While movie metadata is predominantly presented in a structured tabular format, it highlights key attributes such as cast, crew, genre, and release details, which can be readily analyzed using traditional data processing techniques. Furthermore, unstructured data, such as textual movie reviews, and sequential rating data over time offer significant opportunities for advanced analytics. These diverse data types allow for the application of sophisticated methods, such as natural language processing and time-series analysis, to extract valuable, data-driven insights into audience preferences, sentiment trends, and market dynamics.

This study employs advanced machine learning (ML) techniques to investigate the factors driving movie success across all stages of the film lifecycle: preparation, production, publicity, and release. The role of critic and user ratings are evaluated, considering seasonality, star power, and studio size in shaping box office outcomes. Release timing is analyzed for its strategic importance, particularly during high-demand periods, and the unique challenges of sequels are addressed by assessing factors like runtime, engagement, and reviews. Furthermore, natural language processing further complements the research to examine how textual reviews predict ratings, exploring the moderating role sentiment polarity and review length. Finally, multivariate time-series models are applied to predict post-release performance, emphasizing the ability of temporal data and machine learning to optimize decision-making and enhance financial and audience outcomes across the movie industry.

2. Literature Review

2.1. Defining Movie Success

Movie success has a twofold definition depending on the scope of stakeholder. While consumers are more drawn public perception in terms of ratings, movie producers and investors additionally focus on profitability and maximizing financial indicators (Dellarocas et al., 2007; Divakaran, Palmer, Sendergaard, & Matkovskyy, 2017). Therefore, measuring a movie's success involves various factors, including ratings and box office performance emerging as two central success metrics. Ratings, provided by both users and critics, serve as proxies for word-of-mouth, shaping audience perceptions and influencing their viewing decisions. Historically, capturing consumer word-of-mouth was challenging due to its transient nature in offline communities. However, internet-mediated platforms have revolutionized this dynamic by providing a persistent and publicly accessible record of consumer opinions. While these platforms offer valuable insights into audience sentiment, they also highlight the inherent subjectivity of ratings.

Moon, Bergey, & Iacobucci (2010) highlight the complexities of movie ratings, which are influenced by individual preferences, viewing histories, community opinions, and movie characteristics. Another study by Tsao (2014) further underscores the subjective nature of movie ratings, driven by individual biases and diverse preferences. Consumer reviews significantly impact both movie selection and post-viewing evaluations, but the variability in how audiences interpret and respond to positive or negative reviews reveals the inconsistencies in ratings. Factors such as individual taste, mood, and expectations influence ratings, making them a valuable yet subjective measure of success.

In contrast, box office performance provides a more objective and financial measure of success (Divakaran et al., 2017). Box office revenues not only reflect a movie’s profitability but also

influence strategic decisions, such as marketing and distribution. Accurately predicting box office success is critical for optimizing marketing efforts, such as determining the release date, setting the number of screens, and increasing audience awareness through targeted advertising.

Despite their differences, both ratings and box office performance face challenges as success metrics. Movies, as short lifecycle products, have limited windows for revenue generation, making accurate forecasting of early success critical for mitigating financial risks (Chung, Niu, & Sriskandarajah, 2012; C. Zhang, Tian, & Fan, 2022).

These dynamics highlight the importance of strategic planning and predictive modeling in the film industry. By addressing the challenges inherent in both ratings and box office performance, stakeholders can better navigate the risks and uncertainties of the highly competitive movie market, ultimately improving the chances of a film's success.

2.2. Factors Driving Movie Success

2.2.1. User and Critic Ratings in Movie Success

Critic ratings are evaluations provided by individuals or entities with expertise in a specific field (Pang et al., 2022). In cultural industries like movies, books, wine, and restaurants, critics act as intermediaries connecting producers with consumers. They are distinguished by their explicit expertise and authority in evaluating quality and taste within their respective fields.

User ratings are crowd-sourced evaluations shared by general audiences, often through online platforms (Divakaran et al., 2017). In the movie industry, these ratings allow individuals to express their opinions, preferences, and experiences regarding films, including those yet to be released. Movie-based online communities play a vital role in facilitating the exchange of these crowd-sourced opinions.

User ratings are crowd-sourced evaluations shared by general audiences, often through online platforms (Divakaran et al., 2017). In the movie industry, these ratings allow individuals to express

their opinions, preferences, and experiences regarding films, including those yet to be released. Movie-based online communities play a vital role in facilitating the exchange of these crowd-sourced opinions. The movie industry faces significant information asymmetry before a film's release, as consumers have limited knowledge about its quality (Pang et al., 2022). This gap often prompts audiences to consult external evaluations, such as user and critic ratings, to reduce uncertainty. Critics, regarded as experts in their field, connect producers with consumers by providing informed opinions based on their expertise. Similarly, online communities serve as platforms for users to share their opinions and experiences, offering valuable crowd-sourced information (Divakaran et al., 2017). The persistent and publicly accessible nature of online ratings allows potential viewers to form opinions based on the experiences of both peers and experts (Dellarocas et al., 2007).

Studies have explored the relationship between critic and user ratings, revealing a moderate positive correlation between the two (Dellarocas et al., 2007). While there is alignment in some respects, the relatively low correlation highlights the distinctiveness of these sources of information. Critics' evaluations are rooted in professional expertise, yet they may also reflect personal biases or genre-specific preferences (D'astous, Montreal, Touil, & Tunisia, 1999). On the other hand, user ratings are dynamic, changing over time as new viewers contribute their opinions, making them more reflective of audience sentiment throughout a movie's lifecycle (Divakaran & Nørskov, 2016). These differences justify treating critic and user ratings separately, as they provide complementary perspectives (Dellarocas et al., 2007).

Reliance on critic or user ratings depends on the timing and availability of information. When information is limited, such as before or during a movie's release, consumers often turn to critics for their expertise and structured evaluations (Flanagin & Metzger, 2013). However, as more user-

generated content becomes available, consumers shift their reliance toward peer reviews, which are considered more relatable (Divakaran & Nørskov, 2016; Flanagin & Metzger, 2013).

2.2.2. Effect of Electronic Word-of-Mouth (eWOM) in the Movie Industry

Word of Mouth (WOM) and the subcategory of Electronic Word-of-Mouth (eWOM) are common topics of interest in marketing and consumer behavior studies (H. Zhang, Yuan, & Song, 2020). Both address the importance of consumer resumes and feedback for a product's success as it affects product preferences and purchase decisions made by other consumers. Particularly, volume and valence are KPIs of interest within the field of WOM research. They have been demonstrated to impact the success of products and services. For movies, researchers have shown that eWOM can significantly impact a movie's success. Duan, Gu, & Whinston (2008) identified WOM volume (total number of reviews and ratings) as the key driver of the box office revenue of a movie as increased WOM volume. In contrast, WOM's valence (the sentiment of reviews) showed a rather indirect impact on the box office. They also identified a dynamic interaction and positive feedback mechanism between WOM and box office revenue, meaning that the box office revenue can influence the WOM volume, which subsequently can influence the box office revenue again. While their study focused on discussing and investigating the dynamics of WOM and box office revenue, they have yet to delve into how specific patterns can be used to inform detailed marketing strategies of decisions in real time. However, the study of H. Zhang et al. (2020) focused on the role of marketing activity and eWOM in movie diffusion in both pre- and post-release phases. They investigated the role of advertising, WOM, and eWOM on movie diffusion (H. Zhang et al., 2020) by segregating advertising, eWOM, and viewers into two different groups. Users were divided into innovators and imitators, while advertising and eWOM were divided into pre-release and post-release. Their research indicated that pre-release advertising was more effective in attracting the innovator's group, while post-release advertising and positive post-release eWOM were more

influential for imitators. What stood out in their study was that the innovator's group is sensitive to negative pre-release eWOM. In contrast, negative post-release eWOM could increase their interest due to heightened awareness. While their study highlighted the role and effect of eWOM on the movie's performance, it did not focus on the dynamic effect by incorporating temporal data into their model. Moreover, they did not conduct their study model's performance across multiple channels, leaving the difference between different platforms unexplored. Recently, Delre & Luffarelli (2023) found that eWOM significantly impacts the box office revenue, peaking in the early stage of a movie's release and then declining over the cycle of the movies. Mainly, they investigated interaction effects of eWOM volume and valence on the box office sales, indicating that the features show a positive effect during the first 4 weeks after a movie's box office release. They focused on analyzing data from IMDb and Rotten Tomatoes to build a regression analysis that focuses on trends and effects rather than the overall task of modeling success prediction using machine learning based on instance-level temporal and textual features. This indicates that their insights could be used to formulate a predictive model for multivariate TSMC.

2.2.3. Influence of Major Studios on Film Performance

Major studios often act as a quality signal for consumers, as their established reputation and resources suggest a higher likelihood of delivering high-quality productions (Kraft & Rao, 2024). The uncertainty surrounding quality is highest for movies produced by non-major studios. Regression analysis confirms this, showing that returns to signaling are most significant for non-major studio productions compared to major studios and foreign productions.

Major studios like Disney and Warner Bros. are likely to dominate the market through their superior resources, extensive marketing budgets, and robust distribution networks (Pang et al., 2022). According to Basuroy (2006) these elements enhance their ability to signal quality effectively, influencing consumer perceptions and box office performance. If a studio's signals, such as high

upfront investments, are credible, they can positively shape perceived quality and, consequently, box office revenues. Building on these insights, we propose that non-major productions face greater quality uncertainty and are likely to have weaker distribution networks, resulting in less user-generated content about their movies.

2.2.4. Seasonality Effects on Box Office Revenues

Seasonality significantly influences box office performance, with consumer demand varying throughout the year (Einav, 2007; Simonton, 2009). Big-budget films are strategically released during high-demand periods, such as early summer and the Christmas season, to maximize audience reach and revenue. Analysis revealed that movies released during peak seasons often benefit from higher audience turnout, driven by factors such as school holidays, increased leisure time, and reduced work commitments (Ahmad, Duraisamy, Yousef, & Buckles, 2017). In contrast, box office revenues tend to decline during the off-season, particularly from mid-summer to early September . To account for these variations, researchers can include seasonality as a control factor, recognizing that differences in release timing can have substantial effects on a movie's performance (Karniouchina, Carson, Theokary, Rice, & Reilly, 2023).

2.2.5. Strategic Release Timing and Its Impact on Revenue

Numerous studies have identified release timing as a critical determinant of box-office performance. Ryoo, Wang, & Lu (2021) analyzed the impact of release timing across various periods, finding that holiday releases consistently outperformed those during non-holiday windows. Their study attributed this phenomenon to increased audience availability and leisure time during holidays, which boost theater attendance. The authors also highlighted that weekend releases – particularly those combined with holidays – maximize revenues, as they capture both opening-day buzz and sustained attendance over an extended period. Despite these findings, the authors primarily focused on family-oriented films and did not explore the broader implications for

other audience demographics. Future research could build on these findings by systematically comparing timing effects across different timeframes.

2.2.6. Impact of Star Power on Audience Engagement

In this study, star power is measured by the number of followers an actor has on the social media platform Instagram. Actors with significant followings often attract audiences due to their established popularity, creating a strong draw for potential viewers (Divakaran & Nørskov, 2016). This form of social influence contributes to shaping audience expectations and quality perception. User-generated content, such as crowd-sourced ratings, captures a broader spectrum of audience characteristics, including preferences, past experiences, and social influences, compared to critic evaluations(Einav, 2007; Simonton, 2009). While critics emphasize technical and artistic elements, ratings from users reflect mass-market sentiment, making them more representative of general audience opinions. Social influences, such as the popularity of major stars, further enhance the relevance of consumer evaluations, as these ratings often strongly correlate with box office performance.

2.2.7. Brand Extension and Success in Movie Sequels

Sequels play a special role in the context of determining factors influencing movie success. Building on the understanding of general success factors, sequels represent a distinctive case where elements such as brand equity take on heightened importance. As a form of brand extension, sequels not only capitalize on the value of a successful original work but also set expectations for audiences, solidifying their position as reliable box office performers. This dual role underscores their strategic significance in the motion picture industry, where leveraging a well-established brand often translates into consistent audience engagement and profitability (Moon et al., 2010).

The reputation of the original film shapes audience expectations, influencing decisions to engage with the sequel (Belvaux & Mencarelli, 2021) . This reputation is often reinforced through creative

continuity, such as retaining directors, writers, and key cast members. Consistency in the creative team preserves the narrative and tonal coherence of the franchise, strengthening audience trust and delivering on promises established by the original (Moon et al., 2010). Similarly, maintaining genre consistency helps balance audience expectations for familiarity and innovation. Successful franchises frequently reward audiences with storytelling that evolves while adhering to established thematic and stylistic boundaries (Belvaux & Mencarelli, 2021).

Conversely, deviations in creative leadership or genre can disrupt established audience expectations. Changes in directors, writers, or cast members risk breaking the emotional connection audiences feel with the franchise, creating uncertainty about the sequel's quality. Similarly, significant shifts in tone or genre may alienate loyal fans while failing to attract new audiences, leading to what Hennig-Thurau, Houston, & Heitjans (2009) describe as "brand fragmentation." Managing these risks requires a careful balance between continuity and novelty to sustain long-term audience engagement and franchise momentum.

Despite the recognized importance of creative and genre continuity, gaps remain in understanding how these elements interact with broader market forces. Factors such as release timing and audience engagement strategies may amplify or mitigate the impact of brand extension dynamics, yet these interdependencies are underexplored (Lehrer & Xie, 2021). By addressing these gaps, this study seeks to provide a more comprehensive framework for understanding the mechanisms underpinning sequel success within the context of brand extension.

2.2.8. Key Factors Influencing Sequel Performance

Sequel success is shaped by measurable factors such as runtime, release timing, and audience engagement metrics, each contributing uniquely to performance outcomes (Belvaux & Mencarelli, 2021; Moon et al., 2010).

Runtime is a significant predictor of absolute box office performance, often seen as a marker of production investment and narrative depth, which attract audiences and drive revenue (Moon et al., 2010). Empirical studies found a positive relationship between runtime and revenue, with longer films typically generating higher earnings (Lehrer & Xie, 2021). However, excessively long runtimes can alienate some audience segments due to pacing issues or time constraints, underscoring the need for balance (Navarathna, Carr, Lucey, & Matthews, 2019). Despite these challenges, runtime remains a critical factor in shaping audience perceptions and marketability (Lash & Zhao, 2016).

Release timing is another crucial determinant. Shorter intervals between an original film and its sequel help maintain audience interest and preserve emotional connections with the original (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). Conversely, longer gaps risk diminishing engagement, especially in competitive markets with an abundance of new content. On the other hand, excessively short intervals may lead to franchise fatigue, where audiences perceive sequels as rushed or redundant (Moon et al., 2010). Striking the right balance between sustaining interest and avoiding overexposure is crucial for maximizing sequel success.

Audience engagement metrics, such as review volume and sentiment, are also critical predictors. Research shows that engagement volume—reflected in reviews, comments, and ratings—correlates strongly with box office performance, as it signifies heightened audience awareness and interest (Navarathna et al., 2019). While sentiment provides additional context, it typically exhibits weaker correlations with revenue compared to engagement volume (Song, Huang, Tan, & Yu, 2019). These findings underline the importance of audience participation as a driver of success.

Despite the importance of these factors, gaps remain in understanding their interdependencies. For example, the impact of audience engagement may vary depending on release timing, yet such relationships are rarely explored (Lehrer & Xie, 2021). By examining runtime, release timing, and

audience engagement collectively, this study helps to provide a better understanding of the drivers of sequel success and their interrelationships.

2.3. Machine Learning in the Movie Industry

2.3.1. Importance of Accurate Predictions in the Movie Industrie

The prediction of box office performance is a critical challenge in the film industry, as the financial stakes of high-budget productions are significant (Lash & Zhao, 2016). Accurate forecasting allows stakeholders - including studios, investors, and distributors - to allocate resources efficiently, reduce uncertainty, and optimize returns (Lehrer & Xie, 2021). As films continue to grow in production and marketing costs, understanding the drivers of success has become an essential focus for both theoretical research and practical decision-making.

Box office predictions are particularly influenced by a variety of factors, including production budgets, marketing strategies, release timing, star power, and audience engagement (Song et al., 2019). These determinants interact in complex ways, influenced by dynamic market conditions and evolving consumer preferences. Advances in data analytics have significantly improved forecasting models, integrating structured variables (e.g., runtime, genre, and distribution strategy) with unstructured data, such as social media activity and audience sentiment, to provide real-time, actionable insights.

Among the wide array of movie types, sequels occupy a unique position within the industry. Building on the brand equity of their predecessors, sequels often benefit from pre-existing audience loyalty and recognition, making them a popular strategy to mitigate financial risk. This has contributed to sequels consistently ranking among the highest-grossing films globally (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009). However, sequels also present distinctive challenges for prediction. Unlike standalone films, they must balance continuity with novelty - preserving the narrative and thematic elements that resonated in the original while offering fresh

appeal to attract new viewers. Audience expectations, shaped by the first installment, create additional pressure for sequels to perform both critically and commercially, complicating forecasting efforts (Hennig-Thurau et al., 2009).

While many models focus on absolute metrics such as total box office revenue, these fail to capture the nuanced performance measures critical for sequels. Relative success metrics, which assess a sequel's performance compared to its predecessor, provide deeper insights into franchise sustainability and audience retention. For instance, a sequel that generates substantial revenue might still underperform if it fails to exceed the original's success, potentially signaling declining interest in the franchise (Hennig-Thurau et al., 2009).

This comprehensive perspective on predicting box office performance aims to enhance theoretical understanding and inform practical applications in an increasingly globalized and competitive market.

2.3.2. Traditional Machine Learning in the Movie Industry

The least squares method is a foundational tool for analyzing data trends, often visualized as a scatterplot where data points exhibit a roughly linear pattern (Burton, 2021; Fox, 2016). This method identifies the "line of best fit," minimizing the sum of squared deviations between each data point and the fitted line, ensuring the smallest total error and providing a precise linear approximation of the dataset. Building on this principle, Ordinary Least Squares (OLS) regression fits a regression plane to data points with a linear trend, capturing partial relationships between regression coefficients and the dependent variable while considering the influence of other variables (Burton, 2021; Fox, 2016). Although widely used, OLS regression relies on several key assumptions, such as linearity, independence of errors, homoscedasticity, normality of errors, and low multicollinearity among predictors. Violating assumptions can lead to biases, invalid

significance tests, and unreliable predictions, underscoring the importance of verifying these criteria.

OLS regression remains a cornerstone of statistical modeling, offering a robust framework for understanding linear relationships and deriving interpretable results. Its applicability and clarity make it a valuable tool, particularly for initial analyses and straightforward relationships. However, in the context of box office forecasting, where nonlinear interactions and complex predictor relationships frequently arise, complementary methods can provide additional insights. For example, factors such as production budgets, cast popularity, and release timing often exhibit interactions or nonlinear effects that may not be fully captured by linear models (Lash & Zhao, 2016; Somlo, Rajaram, & Ahmadi, 2011). Incorporating advanced techniques, such as machine learning algorithms or hybrid approaches, can help address these challenges while building on the solid foundation provided by OLS (Lash & Zhao, 2016). This combination of methods ensures both interpretability and adaptability to complex datasets, as shown by studies that explore the integration of econometrics and predictive analytics to capture richer patterns in data (Lash & Zhao, 2016; Lehrer & Xie, 2021).

The growing availability of granular data from digital platforms and social media further highlighted the need for more sophisticated methods. Machine learning techniques, such as decision trees, support vector machines, and ensemble methods, have proven effective in addressing these limitations. These methods handle high-dimensional data adeptly, uncovering complex patterns among predictors such as runtime, reviews, release timing, and audience engagement (Lehrer & Xie, 2021; L. Liao & Huang, 2021). Additionally, their ability to incorporate unstructured data, like social media interactions, provides real-time insights into audience behavior—capabilities that traditional econometric models lack (Song et al., 2019).

Hybrid approaches that integrate econometric models with machine learning have emerged as a powerful solution, combining structured and unstructured data to enhance forecasting accuracy. By addressing multicollinearity and leveraging machine learning's flexibility, these models produce predictions that are both precise and theoretically grounded (Lehrer & Xie, 2021). However, many studies remain U.S.-centric, focusing primarily on absolute success metrics and neglecting relative measures that provide deeper insights into franchise sustainability and audience retention (Belvaux & Mencarelli, 2021; Hennig-Thurau et al., 2009).

This study builds on these advancements by employing machine learning and hybrid approaches to predict both absolute and relative success metrics. Incorporating predictors like engagement metrics, runtime, and release timing alongside international data, it seeks to uncover the interdependencies that shape sequel performance. By addressing these gaps, it contributes to a more comprehensive predictive framework, enhancing theoretical understanding and supporting practical decision-making.

2.3.3. Deep Learning Models in the Movie Industry

Deep learning (DL), a specialized subfield of machine learning (ML), marks a major milestone in the advancement of artificial intelligence (AI) technologies (Darwish, Hassanien, & Das, 2020). By utilizing multiple processing layers to model complex and abstract patterns in data, DL has become a foundational component of modern AI applications. Its rapid progress has been fueled by the availability of large datasets, the development of powerful computational architectures, and advancements in algorithms that enable efficient learning at scale (Sastry et al., 2024). These factors have allowed DL to address highly complex problems across a wide range of industries. In the movie industry, DL has made significant contributions, transforming processes such as visual effects creation, audience analytics, and personalized content recommendation (Mühling et al., 2017; Tahmasebi, Ravanmehr, & Mohamadrezaei, 2021; Zhou, Zhang, & Yi, 2019). DL algorithms

enhance professional media production by automating tasks like labeling video content, recognizing faces, and identifying similar visuals, making video analysis and retrieval more efficient and streamlined. DL has also advanced recommender systems that incorporate user behaviors and social influence to enhance the personalization of content. Furthermore, in the film industry, deep learning models that analyze movie posters and other data have shown great success in predicting box-office revenues, helping to reduce financial risks.

Moreover, DL significantly influences related fields like Natural Language Processing (NLP) (Banbhrani, Xu, Soomro, Jain, & Lin, 2022) and Time Series Prediction , which face challenges in processing sequential data due to its complexity and temporal dependencies. By effectively capturing sequential patterns, DL has enabled breakthroughs in these fields, with notable applications extending to the movie industry as well (Y. Liao et al., 2022).

2.4. Natural Language Processing

The rise of digital platforms has led to an overwhelming amount of user-generated content, such as textual movie review. These reviews are a rich source of insights into audience opinions, but the sheer volume makes it challenging to analyze them manually. Natural Language Processing (NLP), a subfield of AI, has become an essential tool for addressing this challenge by enabling computers to understand and interpret human language (Banbhrani et al., 2022). NLP focuses on breaking down and analyzing text, transforming it from unstructured data into a format that machines can process (M. T. Khan et al., 2016).

2.4.1. Sentiment Analysis

One particularly useful application of NLP is sentiment analysis, which identifies the emotional tone of text, whether it's positive, negative, or neutral (Wadawadagi & Pagi, 2020). Sentiment analysis uses machine learning algorithms to process reviews and classify them based on the opinions expressed. This approach is widely explored in the movie industry with many different

applications (Berger, Kim, & Meyer, 2021). Danyal et al. (2024) conducted research focusing on extracting subjective information from film critiques, such as the reviewer's overall sentiment, the film's strengths and weaknesses, and viewing recommendations. Their study utilized sophisticated language models to analyze datasets from IMDD and Rotten Tomatoes. Their results demonstrated that sentiment analysis can effectively discern emotions and attitudes expressed in film critiques, providing valuable insights into subjective opinions and improving the ability to summarize movie reviews. By identifying trends in sentiment, this method helps filmmakers and marketers understand how a movie resonates with its audience. What makes sentiment analysis particularly valuable is its ability to handle large-scale data and uncover insights that might not be immediately obvious. For example, the results from Lee, Jung, & Park (2017) highlighted the feature's importance pinpointing that sentiment is helpful in predicting movie success metrics like ratings or box office success. Overall, sentiment analysis bridges the gap between vast amounts of audience feedback in an unstructured data format. By combining NLP and ML techniques, it offers a scalable, efficient way to understand audience opinions and improve decision-making in the movie industry.

2.4.2. Rating Prediction

Sentiment analysis, while widely useful, tends to oversimplify the nuanced opinions found in complex movie reviews (M. T. Khan et al., 2016). In contrast, rating prediction represents a more sophisticated approach within NLP, as it translates textual feedback into precise numerical ratings, providing a detailed and accurate quantification of user evaluations (Ahmed & Ghabayen, 2022). Star ratings in online reviews provide a standardized measure of perceived quality, with higher scores reflecting greater satisfaction. Converting textual reviews into quantifiable ratings enhances accessibility and interpretability for industry stakeholder, enabling more structured analysis of feedback (Archak, Ghose, & Ipeirotis, 2011). Rating prediction methods are used to predict the

score someone might give to a movie by analyzing the sentiment and content of their written review (Z. Y. Khan, Niu, Sandiwarno, & Prince, 2021). By converting subjective language into standardized satisfaction metrics, these methodologies play a crucial role in quantifying opinions within online reviews.

Recent advancements in machine and deep learning have significantly improved rating prediction by using complex architectures capable of recognizing intricate textual patterns. Deep learning utilizes artificial neural networks with multiple layers to analyze complex patterns in data, often exceeding the capabilities of traditional machine learning techniques (Janiesch, Zschech, & Heinrich, 2021). Each layer plays a specific role, with input layers receiving raw data, hidden layers extracting and transforming features, and output layers delivering the final prediction or result. Generally, the more layers a network has, the deeper it is, which often enhances its ability to learn and achieve better results, especially for complex tasks.

Chambua & Niu (2021) demonstrated the effectiveness of deep learning models in rating prediction, highlighting their ability to detect subtle emotional cues in text. Similarly, Ahmed & Ghabayen (2022) found that a Bidirectional Gated Recurrent Unit (Bi-GRU) model outperformed traditional methods, showcasing the strength of neural architectures in linguistic feature extraction. A Bi-GRU, an enhanced Recurrent Neural Network (RNN), processes data in both forward and backward directions, allowing it to capture contextual information from both past and future words, which improves pattern recognition in NLP tasks (He et al., 2020). Despite advancements in deep learning, challenges in handling domain-specific complexities and limited data have driven interest in innovative approaches like transfer learning.

2.4.2.1. Transformative Approaches: Transfer Learning

The emergence of transfer learning has further advanced NLP by allowing pre-trained models to adapt effectively to specialized tasks with minimal labeled data (Galal, Abdel-Gawad, & Farouk,

2024). Transfer learning models leverage extensive general knowledge gained from large datasets, reducing reliance on large task-specific datasets and improving performance in tasks like rating prediction (Mosin, Samenko, Kozlovskii, Tikhonov, & Yamshchikov, 2023). Fine-tuning is a critical step in transfer learning, where pre-trained models are adapted to the specific requirements of a target task. By leveraging large-scale pre-trained models, which would often be impractical or infeasible to train independently, fine-tuning aligns these models with the nuances of the target task. This approach not only enhances performance but also demonstrates the efficiency of transfer learning in addressing specialized challenges in NLP, enabling the application of powerful models without requiring extensive computational resources or massive task-specific datasets. The foundation for many transfer learning models can be traced back to the introduction of the transformer architecture in the groundbreaking paper *Attention Is All You Need* (Vaswani et al., 2017). This architecture, based on a self-attention mechanism, revolutionized NLP by allowing models to better capture contextual relationships within text. Self-attention helps the model focus on the most important parts of a sequence, such as specific words in a sentence, enabling it to understand context more effectively and make more accurate predictions compared to traditional DL models. Following this breakthrough Bidirectional Encoder Representations from Transformers (BERT) emerged as a state-of-the-art NLP model (Devlin, Chang, Lee, & Toutanova, 2019). BERT's transformer architecture uses bidirectional self-attention, which evaluates each word in a sentence in relation to all others, providing a more comprehensive understanding compared to traditional unidirectional models. BERT's extensive pretraining on datasets like BookCorpus (over 11,000 books) and Wikipedia (2.5 billion words) equips it with deep language knowledge (Rajapaksha, Farahbakhsh, & Crespi, 2021). BERT's strength lies in its ability to understand complex semantic relationships within text, and this can be effectively leveraged by fine-tuning the model for task-specific challenges (Aljrees et al., 2024). Fine-tuning enables BERT

to adapt its extensive pre-trained knowledge to specialized domains, aligning its capabilities with the unique requirements of tasks like movie review rating prediction. By tailoring the model to the specific language patterns and context of the target dataset, fine-tuning maximizes BERT's potential, making it particularly effective for addressing domain-specific applications with high performance (Mosin et al., 2023).

2.4.2.2. Refining Rating Prediction in the Movie Industry

The analyzed literature offers an extensive examination of rating prediction in multiple domains, highlighting the utilization of textual reviews to enhance precision. Ahmed & Ghabayen (2022) examine product and service evaluations from platforms like Amazon and Yelp, illustrating the efficacy of textual data in forecasting ratings for consumer products. Chambua & Niu (2021) additionally provide a comprehensive analysis of rating prediction methods for various consumer products and services. Aljrees et al. (2024) redirect attention to mobile application evaluations from the Google Play Store, highlighting the discrepancies between written reviews and numerical ratings. Although rating prediction in the movie review domain is relatively underexplored, sentiment analysis has been thoroughly examined. Rating prediction has been well explored for products, services, and mobile apps, but remains under-researched in the movie review sector. A key challenge is the high subjectivity of online reviews (Park, Song, & Sela, 2023). Unlike sentiment analysis, which captures general emotions, rating prediction provides a precise, quantitative assessment of nuanced evaluations. Addressing this gap can offer deeper insights into a film's reception for industry stakeholders.

2.5. Time Series Analysis

Similar to NLP, Time series prediction builds upon the sequential nature of data by introducing a temporal dimension, enabling the examination of changes over time (Ismail Fawaz, Forestier, Weber, Idoumghar, & Muller, 2019). Time series can be categorized as univariate or multivariate

sequences. Univariate sequences focus on a single feature, while multivariate sequences incorporate multiple features, which are often referred to as multiple channels (Chen et al., 2024). The scientific domain of time series analysis encompasses two primary tasks: time series forecasting (TSF) (Haben, Voss, & Holderbaum, 2023) and time series classification (TSC) (Ismail Fawaz et al., 2019; Liu et al., 2024). TSF involves predicting the future behavior of variables over specified time intervals, whereas TSC aims to assign labels to samples based on detected temporal patterns. TSF and TSC often involve high-dimensional data that challenges traditional ML methods, highlighting the advantages of deep learning for complex tasks in the movie industry (Y. Liao et al., 2022). For example, DL enables multimodal approaches to predict movie opening weekend revenue or success (Y. Liao et al., 2022; Madongo & Zhongjun, 2023). Similarly, Rhee & Zulkerine (2016) demonstrated how combining movie features like metadata, ratings, and user reviews in neural networks can predict box office profitability. Modeling approaches for time series often rely on foundational neural network architectures, such as recurrent neural networks (RNNs), including long short-term memory (LSTM) models (Ghanbari & Borna, 2021), or graph neural networks (GNNs) (Liu et al., 2024), both of which are typically implemented within a supervised learning framework. However, recent advancements have expanded time series analysis to accommodate big data scenarios where labeled samples are sparse or irregular (Chen et al., 2024). These developments leverage weakly-supervised learning paradigms, such as multiple instance learning (MIL) (Bing & Wang, 2017; Chen et al., 2024; Early et al., 2024; Tibo et al., 2020). Weakly-supervised learning is a machine learning approach where models are trained with incomplete, inexact, or noisy labels, contrasting with the fully labeled datasets typical of supervised learning (Ren, Wang, & Zhang, 2023).

2.5.1. Multiple Instance Learning for Movie Success Prediction

MIL is a deep learning approach that has been successfully applied in various domains. Carbonneau, Cheplygina, Granger, & Gagnon (2018) defined Multiple Instance Learning (MIL) as a form of weakly-supervised learning approach that is structured as a binary problem where the training set is composed of labeled bags, each containing multiple instances, but where labels are assigned to the bags rather than the individual instances. Based on the standard MIL assumption, a bag is considered positive if at least one instance within the bag is positive. For instance, Bing & Wang (2017) applied it for such as Breast Ultrasound Image Classification. Bakdi, Kristensen, & Stakkeland (2022) applied it for an intelligent predictive maintenance system in ship electric propulsion systems. Frameworks such as Early et al. (2024) proposed MIL for time series classification to make temporal dependences more distinguishable. Their approach addressed the issue that supervised TSC methods, often considered as a black-box, lack interpretability, which is why they proposed a novel framework to overcome this limitation and showcased the effectiveness of their weakly-supervised modeling approach via application across multiple domains. According to their writing, the framework offers an inherent mechanism to provide interpretable insights, indicating that the MIL framework can produce interpretable insights for temporal features. Picking up on this work, Chen et al. (2024) proposed a novel MIL model for multivariate time series data, demonstrating significant improvements in model interpretability due to the model's attention layer. This is particularly useful when predicting movie success, as it helps to understand the model's decision-making process. They pointed out how their proposed model could be extended to integrate multi-channel information, possibly by integrating cross-channel temporal attention or positional encoding.

2.5.2. The Role of Sentiment Analysis in MIL

As outlined before, WOM features are useful when it comes to predicting financial KPIs as well as ratings. Review sentiment specifically are crucial drivers to predict ratings and corresponding movie success, increasing sales in the box office. Considering the versatility of MIL approaches across domains, recent works within the field of MIL indicate that the weakly-supervised approach is extendable to fields other than medical appliances or predictive maintenance, such as sentiment analysis of movie reviews. Tibo, Jaeger, Frasconi, & Wrobel (2020) approached the application of movie-related review data in a predictive setting by formulating the sentiment analysis problem based on IMDb movie reviews to construct a nested bag MMIL (multi-multiple instance learning) Classifier. Picking up on this work, Deng & Yiu (2022) proposed a novel pipeline indicating that it is generally possible to integrate textual features extracted from news into a MIL pipeline to forecast stock trends.

2.5.3. Enhancing Model Interpretability in Time Series Prediction

Works in the MIL field, such as Early et al. (2024), indicate a need to explore more sophisticated integration of interpretability techniques in deep learning frameworks. Following Turbé, Bjelogrlic, Lovis, & Mengaldo (2023), interpretability is a critical aspect of machine learning that focuses on making models' decision-making processes understandable to humans. It is especially important in deep learning, where complex architectures often act as black boxes. Assis, Dantas, & Andrade (2024) argue that interpretability enables validation of models, builds trust with stakeholders, and provides insights that can inform practical decision-making when it comes to evaluating the robustness and transparency of predictions made by AI systems. Their results indicate that there is a trade-off between predictor's performance and interpretability (Assis et al., 2024). Looking at the common approaches, existing machine learning benchmarks such as the one for the TodyNet model by Liu et al. (2024) or TimeMIL by Chen et al. (2024) neglect interpretability in their evaluation

when benchmarking their models and aim to provide the highest performance in terms of predictive accuracy. Although Chen et al. (2024) point out their model's ability to deliver interpretable results, they do not extend their study to compare it with other models in terms of interpretability either by visually comparing models using common attribution techniques such as SHAP or Integrated Gradients which are more universally applicable across different architecture types nor by quantifying the results (Turbé et al., 2023).Moreover, Chen et al. (2024) point out that their model is inherently interpretable due to the attention maps produced by the transformer components within the model, the value of attention to produce interpretable outputs is contested in other papers (Bibal et al., 2022; Jain & Wallace, 2019).

Additionally, they do not consider accessing their model's interpretability by applying interpretability metrics such as Infidelity proposed in the Captum framework for interpretability by Kokhlikyan et al. (2020) or other metrics and methods (Namdari & Li, 2019; Yeh et al., 2019).

3. Data

Existing works from Lash & Zhao (2016) showcase the integration of different data sources, suggesting that various movie-related data sources can be combined to build sophisticated predictive solutions. Their framework integrates IMDb and Box Office Mojo data through API calls and web scraping, ensuring comprehensive data acquisition. Madongo & Zhongjun (2023) used different types of movie data to capture deeper insights. They combined this data in a single model to predict how much revenue a movie would make during its opening weekend. Both research papers indicate that movie-related data offers various potentials for analytics and can be processed to leverage various data sources in movie success prediction. Following these insights, chapter three comprehensively outlines the wide range of different datasets and sources used for this thesis. Due to the strong limitation regarding the available resources, the central datasets for Rotten Tomatoes (Leone, 2020), IMDb (Banik, 2017), Twitter (Dooms, 2021), The Numbers (The Numbers, 2024)were extracted from the data science community platform Kaggle and enhanced using additional open-source APIs such as Box Office Mojo or TMDB (Mohamadi, 2024; TMDB, 2023). Furthermore, the existing database was enhanced by social media information from Twitter and Instagram. This chapter briefly introduces each dataset, showcasing their potential for use in movie-related scientific frameworks. The movie industry benefits from a wealth of data that can be acquired in diverse forms and from various sources, offering opportunities for deeper insights and predictive analytics. Structured data, such as tabular formats containing movie metadata (e.g., genres, actors, budgets) and movie ratings, provide well-organized, easily interpretable datasets that form the foundation for traditional machine learning framework. In contrast, semi-structured data, such as multivariate time series (e.g. box office trends over time), captures dynamic patterns and temporal relationships essential for trend forecasting and operational decision-making. These

data types are complemented by unstructured textual data like movie reviews that can be processed using advanced NLP methodologies to valuable qualitative insights into audience feedback and preferences. Given this diverse data environment in the movie industry, this study leverages movie metadata, box office performance data as well as movie ratings and reviews.

3.1. Movie Metadata

Movie metadata appears in a wide variety of different formats ranging from structured to unstructured data. Data was sourced from well-established online movie platforms Rotten Tomatoes and IMDb. Metadata is displayed in a tabular fashion and is considered structured data. It includes various features that can be used to better understand different aspects of a movie. For instance, it contains title, production studio, languages, runtime and information about the cast.

An open-source movie data collection from Rotten Tomatoes offers a detailed repository of metadata for over 140,000 movies, making it a useful resource for exploring the characteristics and performance metrics of films. This dataset provides structured and diverse information that captures both the qualitative and quantitative aspects of movies.

The dataset contains identifying attributes like the movie's title, release year and rotten tomatoes id. The genre column lists thematic categories, often multiple per movie, enabling analysis of genre-specific trends and audience reception. The runtime column details the duration of each movie in minutes. Key creative contributors to a film are also documented. The director field identifies the individual(s) responsible for the movie's overall vision, while the cast column highlights the main actors and actresses, providing a basis for exploring the influence of star power on a movie's success. The data further includes the MPAA rating, which specifies the audience suitability of each film (e.g., G, PG, R). This classification can be examined in relation to the movie's target demographics and reception.

Quantitative metrics in the dataset are particularly valuable for performance analysis. The critics score and audience score columns provide aggregated ratings as percentages, reflecting the reception of the film among professional reviewers and general viewers, respectively. The box office revenue column, expressed in USD, represents the financial success of the movie, enabling research into the economic aspects of the film industry. Additionally, the synopsis column offers a textual summary of the movie, which can serve as a basis for natural language processing applications to analyze thematic content.

The dataset is robust and versatile, but it is not without challenges. Missing values, particularly in fields like box office revenue necessitate careful preprocessing or the use of additional data sources. Similarly, the multi-label nature of the genre field requires appropriate handling to extract meaningful insights.

Additional movie metadata was collected from IMDb. The dataset obtained provides a comprehensive view of cinematic metadata for 45,000 movies released on or before July 2017. This dataset integrates diverse information sourced from the TMDB Open API and GroupLens platforms, offering rich contextual data for movie analytics. It captures a wide range of features, including metadata such as movie titles, production details, budget, revenue, and release dates, as well as artistic elements like language, production companies, and production countries. Additional layers of data include cast and crew information, plot keywords, and user engagement metrics like ratings and vote counts. These features create a robust framework for exploring the interplay between qualitative and quantitative dimensions of cinema, allowing researchers to analyze trends over time, predict performance, and explore audience preferences.

3.2. Box-Office Performance

The box office data provides detailed information on the financial performance of films across domestic and international markets for both their whole box office lifecycle and individual days for a subset of movies. It includes key attributes such as the movie title, worldwide gross revenue, domestic revenue, and foreign revenue. Additionally, the data breaks down the percentage of domestic and foreign markets contributing to the total gross revenue.

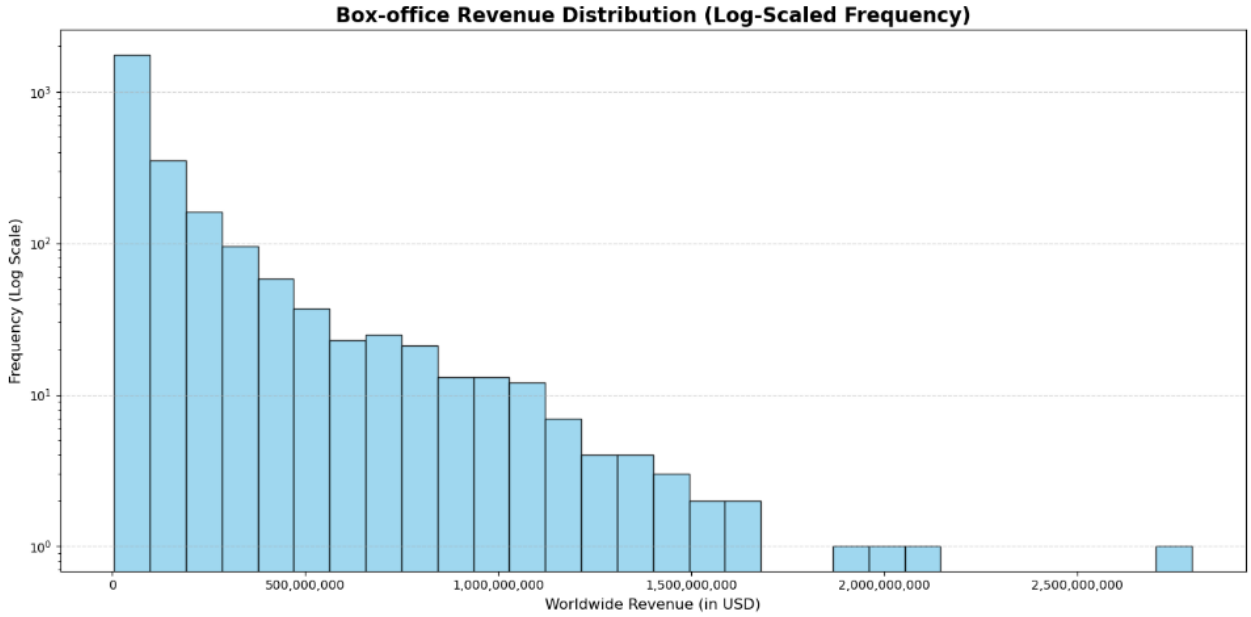


Figure 1: Distribution of Movies Lifetime Box Office Revenues Worldwide

The histogram in Figure 1 displays the distribution of worldwide box-office revenues, with the y-axis scaled logarithmically to better represent the varying frequencies across the available data. The x-axis represents revenue in USD, showing a concentration of films with lower box-office earnings, as indicated by the taller bars near the left. As revenues increase, the frequency decreases exponentially, as seen by the shorter bars toward the right. The logarithmic y-axis highlights the steep decline in frequency, with most films earning significantly less than $500 million, while only a few achieve revenues exceeding $1 billion. This distribution reflects the skewed nature of box-office earnings, where a small number of blockbuster films dominate total revenue.

The data is valuable for examining box-office trends. It allows for cross-referencing with other variables such as release dates, seasonal patterns, and audience behavior to investigate the impact of release timing on financial success. Moreover, the dataset's granularity supports the construction of predictive models to identify the factors that most significantly drive box-office performance.

To analyze the relationship between the original movies' metadata and their impact on sequel box office performance, we required a dataset explicitly identifying sequels. Detecting sequels is not always straightforward, as titles do not consistently include indicators like "Part 2" or similar phrasing. For this purpose, we sourced detailed data from "The Numbers", which specializes in worldwide box office performance for over 1,700 sequels. This dataset provides insights into the financial outcomes of sequel films, a key segment of the movie industry with notable commercial and audience interest.

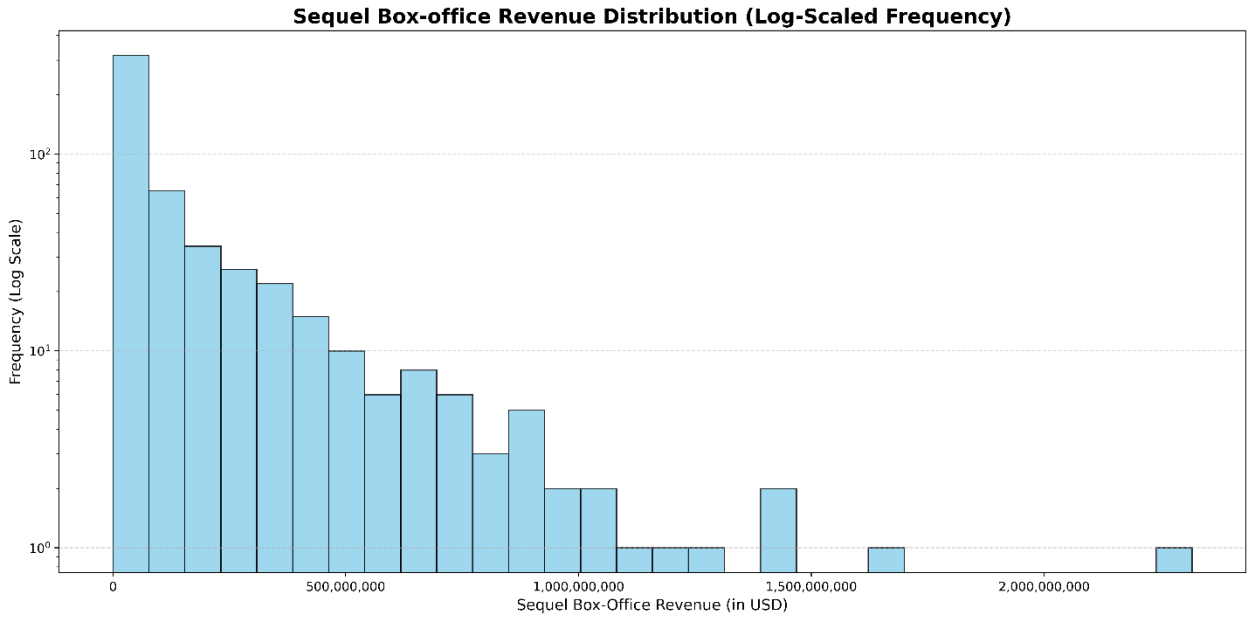


Figure 2: Distribution of Movies Sequels Lifetime Box Office Revenues Worldwide

The dataset includes essential fields such as the sequel title, release year, and cumulative worldwide gross (in USD). These structured metrics enable comparative analyses, helping to identify factors influencing sequel success. For instance, Figure 2 illustrates the distribution of lifetime box office

revenues for movie sequels, showcasing the significant variation in financial performance within this category. Interestingly, when compared to the distribution of general movie box office revenues Figure 1, we observe notable similarities in their patterns. This resemblance suggests that sequels tend to follow the same general economic trends as movies overall. However, this also makes it particularly compelling to investigate what sets sequels apart - specifically, how unique factors like their connection to the original movie or audience expectations influence their performance.

While valuable, this dataset posed integration challenges when combined with other sources like Rotten Tomatoes. The absence of a shared unique identifier (e.g., a universal movie ID) necessitated reliance on title and release year for matching. Variations in formatting and language across datasets further complicated this process. For example, a sequel might be labeled "Part 2" in one dataset and "Part II" in another or titles in different languages (e.g., French or German) often appear inconsistently translated into English.

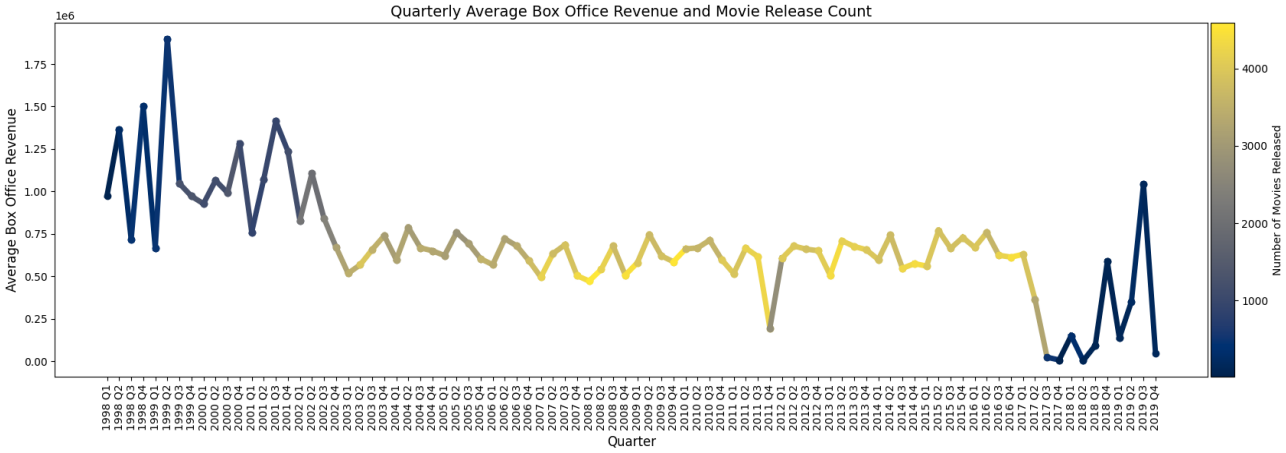


Figure 3: Average quarterly Box Office Revenue recorded across samples in the daily Box Office Revenue dataset

Further, box office data can also be represented in a contextualized time series format, representing the box office more granularly with a timestamp. The time series contains detailed information on daily box office revenues, theater counts, and movie performance metrics as time series, focusing on the USA and Canada markets. Each entry represents a daily record for a specific movie. An

outline of the trends for average quarterly Box Office revenue recorded over the years as well as the number of movies released per quarter of any year is illustrated in Figure 3.

The *daily* column captures the revenue generated by a movie on a particular day, expressed in US Dollars. *Theaters* refers to the number of theaters where the movie was screened on that day, providing insights into its market reach. The *BoxOffice* column aggregates a movie's total revenue. In addition to these key features, the dataset includes various metadata such as release dates, distributors, genres, ratings, and other performance metrics. These ancillary features help contextualize the daily and cumulative box office data. The dataset spans from the years 1999 to 2019 and provides a granular look at movie performance in the targeted regions.

3.3. Movie Reviews and Ratings

Reviews and ratings are very different compared to metadata. Following up on the data analysis, they often appear in a temporal, spatial format and are rather unstructured. Reviews and ratings capture critics and public perception about movies. This feedback-based data provides key-features for examining movie success and therefore provide valuable contributions to predictive analytics within the industry. Several data platforms were conducted to obtain relevant datasets. These platforms capture public perception combining review content, ratings and temporal information.

3.3.1. Movie Reviews: Rotten Tomatoes

Open-source movie review data was obtained from Rotten Tomatoes covering a wide range of critics and user feedback in an unstructured textual format. This data was accompanied by complementary structured features like corresponding ratings and timestamps. With over a million written reviews, the data obtained provides a granular perspective on how diverse audiences react to the movies documented in the accompanying metadata, enabling a detailed understanding of viewing reception and engagement. An exploratory data analysis was conducted to understand the dataset's structure and prepare it for further analysis. Figure 4 displays the distribution of review

length defined as word count in KDE as well as boxplot. The visualizations show that the review content varied significantly, with an average length of approximately 21 words and a standard deviation of 9.42. The shortest reviews contained only a single word, while the longest extended to 55 words. The boxplot further supports highest distribution of word count around the average with a few outliers above the 50 words. The distribution revealed that 75% of reviews were 28 words or fewer, while half of the reviews contained 21 words or less, indicating a predominance of shorter, more concise reviews providing a realistic scenario.

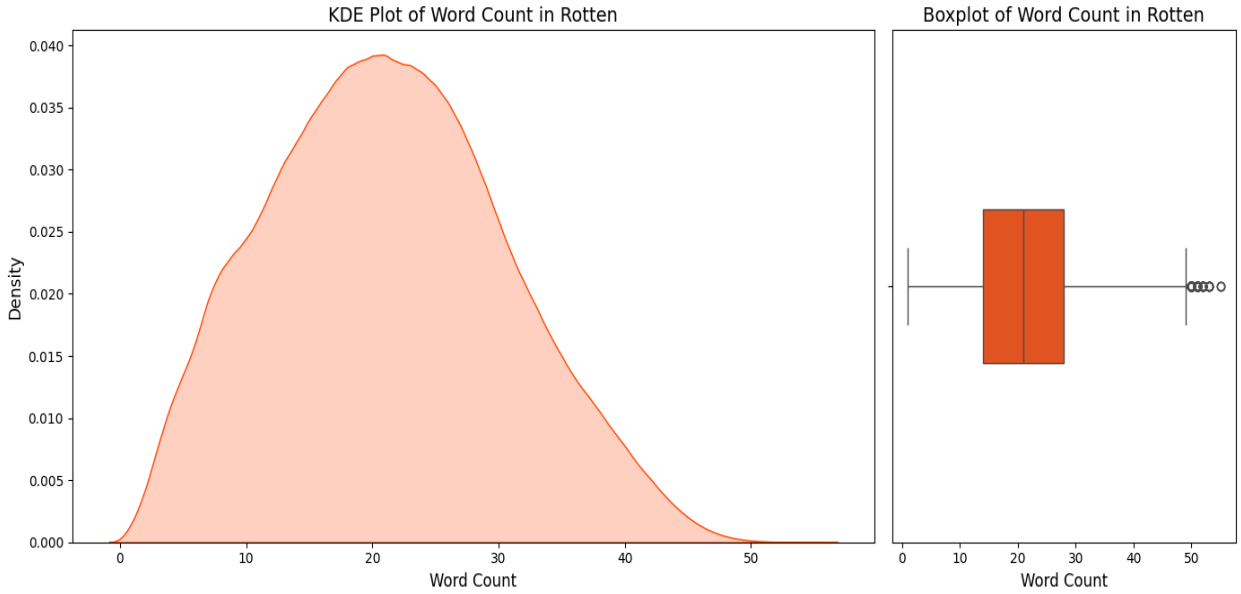


Figure 4: Distribution of Word Counts for Rotten Tomatoes Movie Reviews

3.3.2. Movie Ratings: Perspective across Platforms

Rating data was obtained from various platforms like Rotten Tomatoes as well as IMDb and Twitter. In Rotten Tomatoes data ratings are a critical component of review datasets, providing structured numerical insights to complement the textual content. The dataset obtained included over 1 million reviews on over 17,000 distinct movies. Rating scores were represented in various formats, including numeric fractions (e.g., 3/5), integers on diverse scales (e.g., 7, 100), and letter

grades (e.g., A, B+). Such variability in formats necessitates careful preprocessing to ensure consistency. To enable comparability, all ratings were normalized to a uniform 1–5 scale. Ambiguous data entries, such as "4/5.5," were flagged as errors and excluded from the data. This rigorous normalization process ensured consistency and interpretability, aligning with established best practices in predictive modeling.

Similarly, as the Rotten Tomatoes dataset, the Movie Tweetings dataset created by Dooms (2021) offers a dynamic and contemporary collection of movie ratings sourced from structured tweets on Twitter, encompassing around 39,000 unique movies with around 930,000 ratings recorded between the years 2013 and 2020. Designed to overcome the limitations of static datasets, it leverages real-time social media data, with ratings scaled from 0 to 10 and linked to IMDb identifiers for consistency and metadata enrichment. Twitter ratings are also recorded along with users. This enhances the potential of deriving potential features not only by focusing on the specific rating scores but also on the individual user's perceptions of the specific platform.

While Rotten Tomatoes gives an outline towards professional and semi-professional critic reviews, the IMDb dataset created by (Banik, 2017), specifically captures public opinion for movies as an entertainment product in a more holistic matter in terms of consumer ratings, in a large scale across streaming platforms. The platform captures ratings of viewers at a large scale for millions of different user ratings and reviews across different social media and streaming platforms such as Amazon, Disney+, Netflix, Twitter and many more. As a Big Data service, it captures a more holistic view of public perception and thus, the underlying data distributions can vary in their characteristic. The dataset used within this study captured around 26 million rating samples across 45,000 movies. Similar to the Twitter dataset, IMDb ratings appeared in a more standardized format with rating values ranging between 0 and 10. Moreover, due to the temporal characteristic of the recorded rating, the dataset also lends itself to time-series analyses by providing temporal

information on release dates and user interactions, which can be used to explore historical trends in cinema. The combination of metadata and audience-driven insights offers an opportunity to contextualize movie success and cultural impact within a broader temporal and industrial framework, making it a versatile resource for studying the evolution of the film industry and its audience dynamics.

3.3.2.1. Rating Distribution across Platforms

Looking at the distribution of ratings across platforms displayed in Figure 5, various insights can be overserved. Ratings for Rotten and IMDb both exhibit a slightly left-skewed pattern, with Rotten centering around a score of 3 and IMDb around 3 to 4, reflecting a balance of positive and negative feedback. In contrast, Twitter is heavily right skewed, with the majority of ratings clustered at 4 and 5, indicating a strong positive sentiment bias. The disparity underscores the distinct purposes of these platforms: Rotten Tomatoes and IMDb are established as critical platforms trusted in the entertainment industry, incorporating a combination of user and critic feedback, which results in more balanced distributions. Meanwhile, Twitter, as a multipurpose social platform, encourages informal and spontaneous sharing of opinions, often leading to an emphasis on enthusiastic and positive sentiment at least in the available samples in the dataset used for this word.

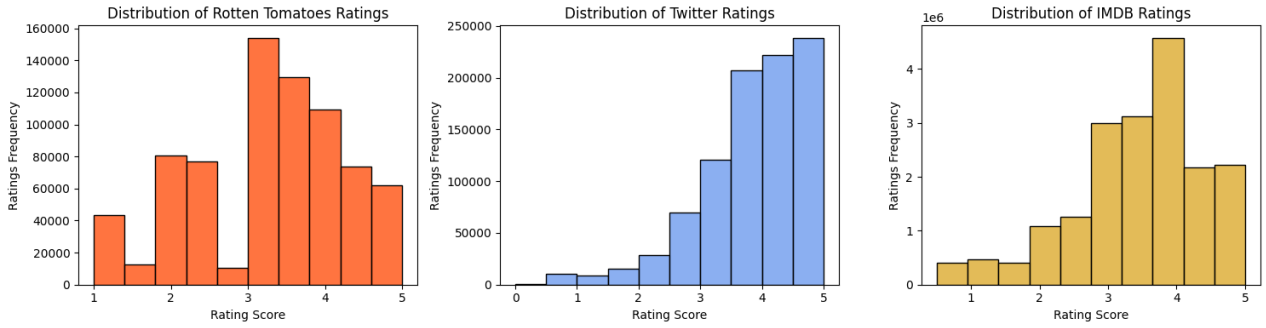


Figure 5: Rating Distribution across Rotten Tomatoes, Twitter and IMDb

3.3.2.2. Ratings in a Temporal Sequential Perspective

As mentioned, ratings and reviews are mostly attributed to a timestamp, extending them to time series. Following this, each movie's ratings form a sequence of values evolving in time, enabling spatial-temporal analysis to identify timely patterns that can be attributed to certain features within the data. Typical patterns are trends or seasonality, as outlined in section 2.2.5. An important characteristic of the data across different platforms is the varying length of the time sequences across features. Figure 6 visualizes the different distributions of time sequence lengths per movies across the three datasets. A filter was applied to include movies recorded for equal to or more than 28 days. However, it is to be mentioned that, due to movies differing in terms of time spent at the box office, samples can range from single days to years. The sequences for each platform show unique characteristics as each platform captures different perspectives within the market. For instance, the data derived from Rotten Tomatoes captures professional and non-professional critic reviews that often occur before and after the official release of a movie. IMDB and Twitter, on the other hand, capture the opinion of the mass audience in the post-release market stage, resulting in more significant amounts of available samples and, thus, longer time sequences. The broad span of unique characteristics within the data and features across platforms offers the opportunity for multifaceted analysis, combining different methods as well as data engineering techniques to form unique feature-sets for advanced data analysis within the movie industry.

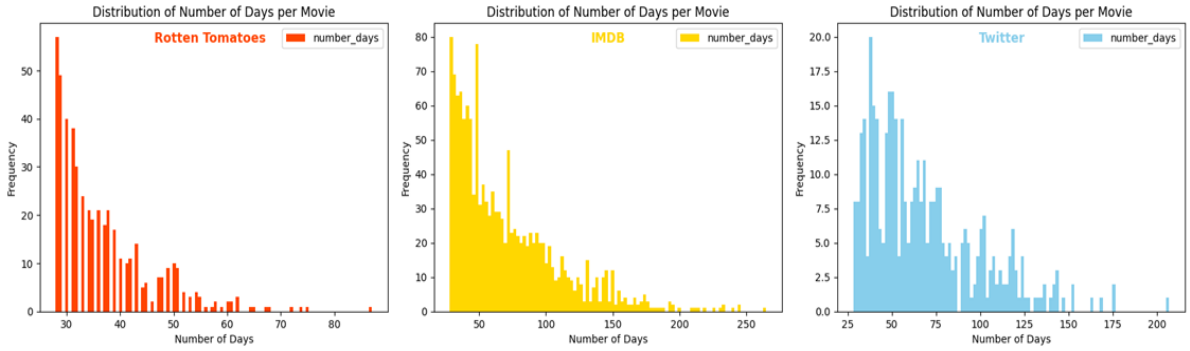


Figure 6: Distribution of Sequence Lengths per dataset.

4. Introduction to Individual Parts

Based on various data types and sources introduced in the previous section of this research paper, a comprehensive machine-learning framework exploring various factors that drive movie success is proposed.

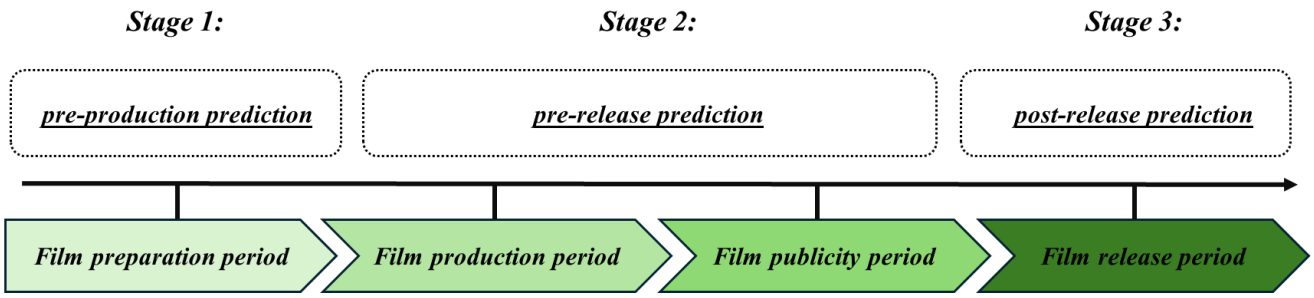


Figure 7: Movie Lifecycle showcasing three stages for Machine Learning opportunities

Figure 7 summarizes the main stages of a movie production lifecycle, defining characteristics within this diverse domain (Y. Liao et al., 2022). The cycle is divided into four sequential periods: Film Preparation Period, Film Production Period, Film Publicity Period, and Film Release Period. Each phase represents a critical stage in the filmmaking and distribution process. The preparation period involves pre-production activities, such as script development, casting, and budgeting. The production period focuses on filming and post-production tasks, including editing and visual effects. The publicity period encompasses promotional activities designed to generate awareness and excitement about the film. Finally, the release period marks the film's debut to audiences through theatrical or digital distribution channels, targeting revenue generation and audience engagement. This timeline highlights the structured progression of activities essential to a film's success. It also enables the directed deployment of various machine learning frameworks and predictive modeling approaches to enable informed decision making at various stages of a movie lifecycle.

This study examines the application of comprehensive machine learning frameworks across key stages of the film lifecycle, focusing on preproduction, prerelease, and post-release predictions. The post-release phase emphasizes feedback from ratings, reviews, and box office performance to evaluate success. The prerelease phase explores factors such as release timing and post-production elements to optimize audience engagement. Additionally, pre-production prediction is addressed by exploring actor cast as well as sequel production. By addressing these phases, the study highlights how predictive modeling supports data-driven decision-making throughout the film lifecycle. The following individual part is introduced:

Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics

With focus on multivariate time series classification, this study explores the applicability of Multiple Instance Learning (MIL) models for predicting box office performance based on time series data in the post-release stage of a movie. In a comparative approach, an MIL model was tested against supervised models for predictive power and interpretability. The research sheds light on the strengths and limitations of weakly-supervised approaches, particularly in handling movie-related features such as time series data and gives implications on how future model benchmarks could be conducted in a more use case-ready way.

5. Temporal Modeling for Movie Success Prediction: A Comparative Study on the applicability of a MIL Model in Entertainment Analytics

5.1. Introduction

Existing research has extensively examined the characteristics and importance of various input features in identifying drivers of movie success, with a strong emphasis on model and feature interpretability (Lee et al., 2017). Features such as review sentiment and audience ratings have been consistently analyzed to understand their impact on box office revenues (Duan et al., 2008; Lee et al., 2017). Temporal patterns, including seasonality effects, have also been identified as critical factors influencing box office performance (Einav, 2007; Karniouchina et al., 2023; Simonton, 2009). Building on this body of work, this study shifts focus to emphasize the temporal characteristics of features within the available data. It explores the application of advanced deep learning frameworks for predicting box office success in the post-release stage. By integrating temporal and post-release features as multivariate time series, this research seeks to explore the accuracy and interpretability capabilities of the weakly-supervised TimeMIL (Chen et al., 2024) compared to a long-short-term recurrent neural network (LSTM) (Ghanbari & Borna, 2021) as a baseline and TodyNet (Liu et al., 2024) representing an advanced supervised architecture.

5.2. Research Question & Hypothesis Development

Deep learning (DL) models for movie success predictions have explored advanced techniques, including multimodal frameworks (Madongo & Zhongjun, 2023) and stacking fusion methods (Y. Liao et al., 2022). While Y. Liao et al. (2022) argue that post-release success prediction has limited value for initial investment decisions, other studies highlight the benefits of incorporating external word-of-mouth (eWOM), user sentiment, and ratings from social media and consumer platforms, which can significantly improve predictive accuracy (Rhee & Zulkerine, 2016;

Subramaniyaswamy, Viginesh, Vishnu, & Logesh, 2017). Although post-release forecasts may not guide initial funding strategies, they offer insights for marketing and distribution planning in the post-release stage (C. Zhang, Tian, & Fan, 2022). This underscores the need for interpretable methods that identify key performance drivers. Multiple-instance learning (MIL), a DL framework that assigns labels to sets of instances (bags) rather than individual samples (Carbonneau et al., 2018), has shown promise in improving interpretability, as evidenced by studies in other domains (Early et al., 2024). While MIL has been used in other fields to integrate textual and temporal features, including sentiment data, for financial forecasting or classification (Deng & Yiu, 2022; Tibo et al., 2020), its application to movie performance prediction remains underexplored. Chen et al. (2024) introduced TimeMIL, a foundational weakly-supervised MIL model for multivariate time series classification (MTSC) and demonstrated competitive performance against supervised models on UEA benchmark datasets (Bagnall et al., 2018; Middlehurst et al., 2024). However, their focus on standard benchmarks does not address real-world use cases and characteristics of the movie domain. Whether MIL models can outperform supervised methods in predicting movie success and delivering interpretable outputs has yet to be established. While Chen et al. (2024) highlighted the inherent interpretability of MIL due to its attention mechanism, concerns raised by Jain & Wallace (2019) and Bibal et al. (2022) about the reliability of attention-based explanations suggest the need for further validation. This study addresses these gaps through the following research question (RQ): To which extend can a weakly-supervised MIL model integrating early movie performance data and eWOM sentiment analysis accurately classify movies into success categories (flop, average, hit) and provide superior performance and interpretability compared to traditional time-series classification methods? To explore this question, several hypotheses are proposed:

H1*: Weakly-supervised MIL models for time series, utilizing early movie performance data (e.g., critic reviews, audience ratings, and box office revenue), will outperform supervised time-series classification methods in predicting movie success.* The hypothesis is supported by prior works demonstrating MIL's efficacy in leveraging textual and temporal features (Deng & Yiu, 2022; Tibo et al., 2020) and Chen et al. (2024)`s findings on MIL's competitive accuracy to other supervised models such as TodyNet (Liu et al., 2024).

H2: *Weakly-supervised MIL models for time series will offer more interpretable insights into factors contributing to a movie's success compared to supervised methods, as evaluated through attribution techniques.* H2 addresses the need for pipelines interpretability, building on research on feature-contributions for predictive models in the movie industry (Y. Liao et al., 2022; Yang, Xu, & Tu, 2023). Prior studies further underscore the importance of identifying key factors influencing movie performance, validating the need for model-specific feature interpretations (Rhee & Zulkerine, 2016; Subramaniyaswamy, Viginesh, et al., 2017). It is supported by recent work on the advantageous value of the MIL framework for interpretability (Early et al., 2024). TimeMIL represents a novel framework to perform multivariate time series classification in the MIL framework. However, its applicability is not yet widely evaluated for the specific use cases of movie success prediction. Consequently, this study proposes to explore the advantages of the MIL framework by comparing TimeMIL to supervised models for the use case of movie success prediction. It will incorporate ratings, sentiment analysis, and box office revenue into time-series data, to address accuracy and interpretability in movie success predictions.

5.3. Methodology

This section outlines the data sources, preprocessing steps for creating the final time series datasets, and the framework for comparing the models in terms of interpretability and performance. All datasets were curated, processed, and combined to create comprehensive time series datasets based

on daily ratings and critical sentiments of movies. This enables the reimplementation and exploration of predictive models for movie success classification based on temporal input features. Figure 8 visualizes the experiment pipeline to outline the different steps conducted in the study.

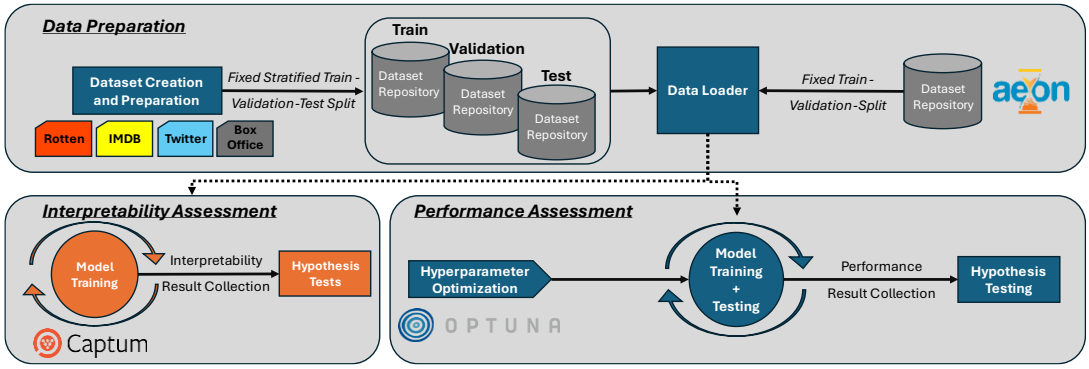


Figure 8: Evaluation Framework to assess the Performance and Interpretability of Models

5.3.1. Datasets and Data Preparation

The central datasets for this study consisted of rating data retrieved from Rotten Tomatoes (Leone, 2020), IMDb (Banik, 2017), as well as Twitter (Dooms, 2021). Additionally, the time series was enhanced by adding movies lifetime revenues, budgets as well as daily Box Office revenues for each movie using data from Box Office Mojo (Mohamadi, 2024). The data, as outlined in section 3.3, specifically the ratings from Rotten Tomatoes, were standardized and cleaned. Further, missing values for budgets and revenues were imputed by integrating APIs such as TMDB during preprocessing (Lash & Zhao, 2016). To form the final time series, the features were aggregated for each platform by calculating mean, min, max, and variance scores for the platforms' rating features to create a time series over daily observations for each movie. As critic reviews were available for the Rotten Tomatoes dataset, the critics' sentiment was extracted and combined with daily statistics like the rating feature (Yang et al., 2023) using the most promising approach outlined in the preceding results of this work focusing on sentiment polarity feature engineering. The data was then joined with daily box office performance data, including daily box office revenues, the daily number of theaters where the movie was shown, and daily mean box office gross per theater. Table

5 and Table 6 in the appendix summarize the features and label distributions over each dataset. Target labels were derived for each movie by estimating the movie's return-on-invest (RoI) using the movie's budget and lifetime revenue described in Equation 1 in the appendix. As outlined by Rhee & Zulkerine (2016), following a classification approach to label movies into three distinct classes can lead to accurate predictions. Moreover, by producing both class labels and associated probabilities, the model reveals how close a film is to flopping, rather than locking it into a single rigid category. Consequently, three classes offer a good trade-off for task complexity for the model while giving enough insight to derive a particular trend when looking at the output probabilities. Thus, the RoI was encoded with the label's "flop" (RoI > 1.0), "average" (1.0 < RoI < 3.0), and "hit" (3.0 < RoI). As described in section 3.3.2.2 movies had between 1 and 270 recorded samples, requiring sequence length adjustments during preprocessing. To ensure sufficient data, only movies with over 28 samples post-release were included, with sequences padded or truncated to a maximum of 60 days fitting them to the median sequence lengths across the different datasets. Additionally, the data was standardized for the training process due to varying ranges between features (Fox, 2016). Following the standard practice and to ensure comparability, each model was trained on the same stratified train-validation-test split (Chen et al., 2024; García, Herrera, & Es, 2008; Liu et al., 2024).

5.3.2. Model Selection

During the experimentation, three models were compared from three different domains. Particularly, the weakly-supervised TimeMIL model proposed by (Chen et al., 2024) was reimplemented in order to compare its performance along with TodyNet, a supervised GNN model proposed by Liu et al. (2024) that showed competitive performance for at least 14 out of the 26 UEA datasets (Bagnall et al., 2018) in the benchmark paper of Chen et al. (2024). Each of the models captures distinct characteristics of different architectural concepts. On the one hand,

TimeMIL, an architecture based on the MIL framework, uses a tokenized transformer with wavelet-based positional encoding to extract sparse and temporal patterns. Wavelet positional encoding applies transformations to capture hierarchical, positional information from sequential inputs. The extracted patterns are then learned by employing a weakly-supervised learning strategy (Chen et al., 2024). TodyNet, on the other hand, employs the framework of dynamic graph neural networks to capture spatial-temporal dependencies through dynamic graph construction and hierarchical pooling (Liu et al., 2024). Additionally, to explore how TimeMIL and TodyNet compare in terms of performance, a baseline model was implemented to capture whether any of the models generally perform significantly differently compared to less sophisticated architectures. To establish a baseline, the LSTM architecture is deployed, as it is considered a well-established DL technique for time-series predictions. LSTMs focus on modeling long-term sequential dependencies with memory cells, differing from the other models' graph-based and weakly-supervised approaches (Ghanbari & Borna, 2021). TodyNet and TimeMIL had to be reimplemented using the available information from their corresponding GitHub repositories (Chen et al., 2024; Liu et al., 2024).

5.3.3. Assessment of Model Performance

To ensure the correct reimplementation of TimeMIL and TodyNet, they were tested on UEA datasets (Bagnall et al., 2018) via the Aeon package by Middlehurst et al. (2024). Additionally, to maintain a fair comparison between the three architectures, each model's hyperparameters were tuned for 30 trials using the Optuna package (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). This reduced randomness when choosing optimal hyperparameters for the comparative experiments. Finally, to evaluate model performance, each fine-tuned model was then trained 30 times for 200 epochs minimizing CrossEntropyLoss (Ho & Wookey, 2020) using the AdamW-optimizer (Loshchilov & Hutter, 2017) together with a ReduceLROnPlateau-scheduler (Al-Kababji, Bensaali, & Dakua, 2022) with a patience level of 50. Regarding the selection of performance metrics,

Accuracy, F1, Recall, Precision, and AUC were calculated during each training run. Due to class imbalances, and as Accuracy tends to hide classification errors while not accounting for the class distributions (Grandini, Bagli, & Visani, 2020), the focus of the performance comparison was the macro average F1 and macro average AUROC score. They reflect performance across all classes more fairly by integrating precision-recall trade-offs and overall discriminative capability, leading to a better estimate of the real performance of each model.

5.3.4. Assessment of Model Interpretability

When it comes to interpretability techniques, three different categories can be distinguished. Some techniques propagate sample and feature importance via model gradients, which is why they are referred to as gradient methods (Schlegel & Keim, 2021). Common gradient-based methods are Saliency (Simonyan, Vedaldi, & Zisserman, 2013), Integrated Gradients (IntGrad) (Sundararajan, Taly, & Yan, 2017), or Deeplift (Shrikumar, Greenside, & Kundaje, 2017). Other techniques use network weights and biases together with a per-model layer-defined set of weighing rules. Furthermore, more sophisticated techniques include surrogate methods such as SHAP, Shapley, or Lime, which generate interpretable attributions by sampling and modeling perturbed instances (Schlegel & Keim, 2021). To compare the models in terms of interpretability, the attributions were evaluated using the metric Infidelity. Infidelity measures how faithfully the applied techniques and attributions capture the model's internal reasoning and assesses how attributions to a model align with the changes in its output when the input is perturbed, reflecting the explanation's accuracy, which is why it was chosen as the primary metric. Perturbation is an important concept in attribution evaluation by applying small changes to the model's input data to see how the model adjusts. Low infidelity scores indicate that attributions align well with the model's behavior when the input is minimally (Kokhlikyan et al., 2020; Yeh et al., 2019). Following Schlegel & Keim (2021), visually evaluating models' feature importance using attributions is still a difficult task even

for domain experts. Thus, to further investigate the visual dispersion and certainty of attributions, the entropy was calculated for the model's attributions to estimate how spread out the assigned importance is across all input features. A low entropy value indicates that the model's attributions focus on several key features and time steps. Conversely, a high entropy indicates that a model's importance scores are more evenly distributed, making it harder to pinpoint which features drive its predictions (Namdari & Li, 2019; Perez, Skalski, Barns-Graham, Wong, & Sutton, 2021).

5.3.5. Hypothesis Assessment

To ensure scientifically correct hypothesis test application, the distribution of the metrics was estimated by applying the Shapiro-Wilk test on the pairwise differences of evaluation metrics between the MIL and the supervised models (García et al., 2008; Shatz, 2024). Depending on the distribution of the performance metrics, the MIL model was compared in a pairwise approach either by applying a paired t-test as the primary method or a Wilcoxon signed rank test as the secondary method (Rainio, Teuho, & Klén, 2024). Additionally, the Bonferroni correction was applied to standardize the p-values across tests. In case the results of these tests are ambiguous, a more novel significance test, referred to as the Almost Stochastic Order (ASO) test (del Barrio, Cuesta-Albertos, & Matrán, 2017; Dror, Shlomov, & Reichart, 2019), implemented by Ulmer, Hardmeier, & Frellsen (2022), was applied to investigate stochastic dominance and assess the hypothesis stated in this study. The test calculates epsilon ϵ so that H_0: $\epsilon_{min} \geq \tau$ can be tested. Epsilon can be interpreted as a confidence score similar to a p-value but does not represent the same. The lower it is, the certain one can be that the model is better than the model it is compared to. Similarly to standard statistical tests, the boundary tau (τ) (commonly set to a value of 0.5 or 0.25) can be interpreted similarly to the significance boundary alpha (α) in standard tests (Ulmer et al., 2022).

5.4. Results

5.4.1. Model Performance Evaluation

Table 1 summarizes the model's average F1 Scores and AUROC values over multiple runs for the different datasets. TodyNet consistently achieved the highest F1 Score for the Rotten Tomatoes and IMDb datasets, with scores of 0.510 and 0.553, respectively. The LSTM performed nearly as well as TodyNet for IMDb, achieving a close F1 Score of 0.552. However, for the Twitter dataset, TimeMIL outperformed the other models with an average F1 Score of 0.504, while TodyNet and LSTM trailed behind with scores of 0.491 and 0.424, respectively across experiments. These results indicate that while TodyNet generally excels in classification tasks across datasets, TimeMIL demonstrates a comparative advantage when predicting data from Twitter. Regarding the models' ability to distinguish classes, similar to the F1 Score, TodyNet consistently achieved the highest AUROC for the Rotten Tomatoes and IMDb datasets, with AUROC values of 0.749 and 0.764, respectively. Similarly, TimeMIL excelled, although by a very minor proportion, on the Twitter dataset, compared to TodyNet and LSTM. However, TimeMIL, while outperforming LSTM in the Twitter dataset for AUROC and F1 Score, generally had the lowest AUROC scores across the other datasets, indicating relatively weaker overall discriminability. Following the results of the Shapiro-Wilk test summarized in Table 12 in the appendix, all the pairwise differences were normally distributed, so the paired t-test was applicable to evaluate the model's significant performance.

Table 1: Comparison of models' average performance on the F1 Score and AUROC

Models	***TimeMIL***		***TodyNet***		***LSTM***		***Rank TimeMIL***	
Datasets\Metrics	***F1***	***AUROC***	***F1***	***AUROC***	***F1***	***AUROC***	***Abs F1***	***Abs AUROC***
Rotten Tomatoes	0.456348	0.658596	**0.510062**	**0.749291**	0.503833	0.698412	3	3
IMDb	0.459950	0.651499	**0.553272**	**0.763592**	0.552268	0.742273	3	3
Twitter	**0.503830**	**0.750273**	0.491021	0.750155	0.423803	0.729751	1	1

Table 2 summarizes the hypothesis test results for pairwise comparisons with a significance level alpha (α) of 0.05. Based on the paired t-test results, significant differences were observed for most

pairwise comparisons over each dataset and evaluation metric. The hypothesis testing results validate the observations for the average performance of the models for the Rotten Tomatoes and IMDb datasets. They confirm that TodyNet excels for movie review data as time series and the GNNs competitiveness for the Twitter time series. Although TimeMIL is superior to the LSTM model regarding the Twitter dataset, its performance is not statistically superior to TodyNet. There was no significant difference between TimeMIl and TodyNet for both the F1 score and AUROC. This indicates that, although the global mean comparison indicated minor superiority of TimeMIL to TodyNet, the performance was not significant enough across experiments.

Table 2: Paired t-test for statistical significance of models superiority across evaluation metrics

MIL-Model	***Metric***	***Dataset***	***Test statistic***	***p-value***	***Test Result***
TimeMIL vs. LSTM	*F1*	*Rotten*	*4.5676*	*0.000084*	*A significant diff. for LSTM*
		IMDb	*14.3439*	*<0.001*	*A significant diff. for LSTM*
		Twitter	*-6.6023*	*<0.001*	*A significant diff. for TimeMIL*
	AUROC	*Rotten*	*5.2239*	*0.000014*	*A significant diff. for LSTM*
		IMDb	*15.3737*	*<0.001*	*A significant diff. for LSTM*
		Twitter	*-3.2751*	*0.002737*	*A significant diff. for TimeMIL*
TimeMIL vs. TodyNet	*F1*	*Rotten*	*5.0527*	*0.000022*	*A significant diff. for TodyNet*
		IMDb	*7.6377*	*<0.001*	*A significant diff. for TodyNet*
		Twitter	*-1.03*	*0.311506*	*No significant diff. for TimeMIL*
	AUROC	*Rotten*	*11.7481*	*<0.001*	*A significant diff. for TodyNet*
		IMDb	*16.6109*	*<0.001*	*A significant diff. for TodyNet*
		Twitter	*-0.0185*	*0.98533*	*No significant diff. for TimeMIL*

5.4.2. Interpretability Evaluation

The interpretability assessment involved attribution techniques such as such as IntGrad or DeepLift. According to the results presented by Nguyen & Martínez (2020), Integrated Gradients provide the most general explanations while offering low effective complexity, meaning that the attributions are efficient and simple to interpret. Additionally, Deeplift was used to ensure comparability over several interpretability techniques. IntGrad and Deeplift ranked in the top tier in the assessment of attribution techniques for the transformer model and bidirectional LSTM model in the comparison of attribution techniques of Turbé et al. (2023). SHAP seemed to be a more sophisticated technique but proved unfeasible due to the incompatibility of SHAP with certain

layers in the compared networks, such as LayerNorm or Identity from TimeMIL. Similarly, although ranked highest in the assessment of Turbé, Bjelogrlic, Lovis, & Mengaldo, Shapley values calculation was not applicable, which can be explained following the findings of Kolpaczki, Bengs, Muschalik, & Hüllermeier (2024), due to the number of features leading to extensive computational complexity of Shapley values resulting in $O(n*n!)$ complexity with n as the number of features. Figure 9 visualizes an example of how attributions of methods such as DeepLift can be used to interpret how a sample is evaluated by the model when classifying the sample (movie) and which features and timestamps contributed negatively, neutrally, or positively to the decision. The red color signals that the timestep had a negative contribution, while the green signals positive attribution scores for the step. Although this can be rather abstract to interpret, such attributions can still be used to determine whether certain timesteps and features had a more substantial influence during the classification process of the sample movie than other time steps. The accuracy of model attributions can be evaluated by calculating the infidelity score for each model. Table 3 summarizes the average infidelity for each model, attribution technique, and dataset.

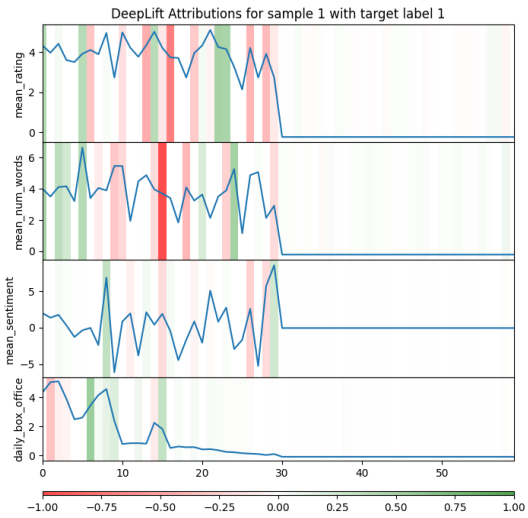


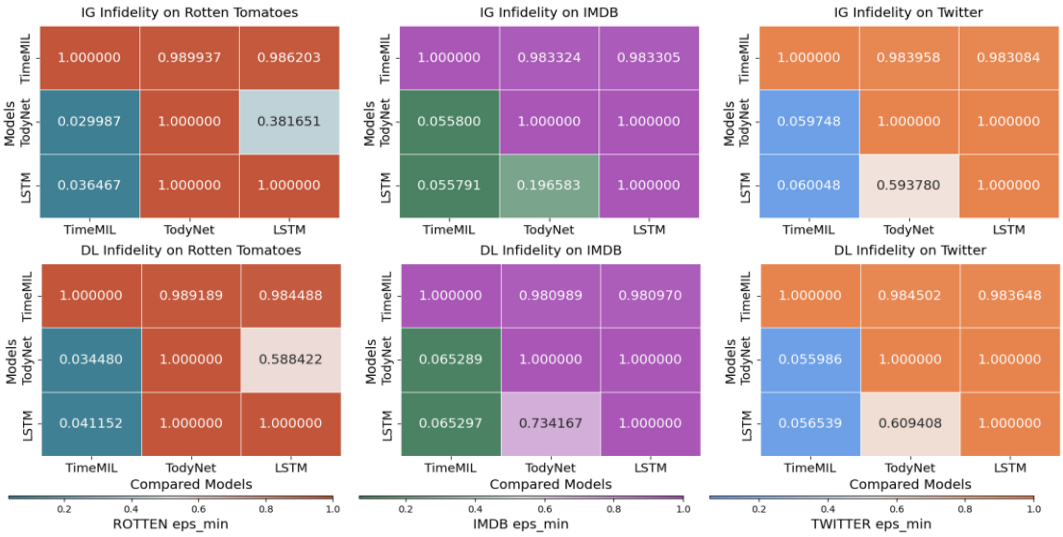
Figure 9: Example of Deeplift attributions on a movie from the test sample labeled as 'average'

Table 3: Comparison of interpretability performance on the experiments average Infidelity across models

Models	***TimeMIL***		***TodyNet***		***LSTM***		***Rank TimeMIL***	
Dataset\Infidelity	***IntGrad***	***DeepLift***	***IntGrad***	***DeepLift***	***IntGrad***	***DeepLift***	***IntGrad***	***DeepLift***
Rotten Tomatoes	0.003376	0.003767	**0.000529**	**0.000514**	0.000781	0.000741	3	3
IMDb	0.573625	0.773938	0.001446	0.001152	**0.001098**	**0.001143**	3	3
Twitter	0.022442	0.025024	0.000584	0.000522	**0.000384**	**0.000341**	3	3

When comparing the average entropy scores in Table 13 in the appendix to the infidelity scores for TimeMIL, one can observe that the average Entropy scores for TimeMIL are larger compared to TodyNet and the LSTM, indicating that its attributions are more evenly distributed. During the assessment of the infidelity metric, the traditional hypothesis test resulted in ambiguous statistics and p-values, making a conventional hypothesis testing approach inapplicable to evaluating the infidelity metric. The statistical tests results for the Shapiro Wilk and Wilcoxon test are summarized in Table 14 and Table 15 in the appendix. Thus, the multiple ASO test with a confidence level of 0.95 was applied to estimate the significance of the model's interpretability performance. Table 4 summarizes epsilon ϵ from the ASO test for the infidelity metric across the attribution schemes and models. The model in the row is stochastically dominant over the model in the column if H_0 can be rejected. Considering $\tau = 0.2$, TodyNet and the LSTM showcase stochastic dominance over TimeMIL, and H_0 for both models can be rejected over all three datasets when compared to TimeMIL. Regarding Deeplift, similar results indicate that the models produce more faithful attributions when used with the methodology compared to TimeMIL. Noticeably, for the Rotten Tomatoes dataset, ϵ for IntGrad and Deeplift is lower for the supervised models, indicating that the technique is stochastically even more dominant compared IMDb and Twitter data.

Table 4: ASO Epsilon Scores across Attributions Methods and Datasets for each architecture



5.5. Discussion

Although Assis et al. (2024) argued that there is a trade-off between performance and interpretability, this characteristic was not reflected in this study's results when comparing weakly supervised to supervised models for predicting movie success via temporal modeling. The supervised models performed superior across evaluation dimensions for performance and interpretability, significantly outperforming the weakly-supervised TimeMIL model. Building on the work of on selecting the optimal evaluation metric, one can see why accuracy is not always the best metric to rely on when evaluating and comparing models (Grandini et al., 2020). This is undermarked by the results summarized in the box plots for the metrics Accuracy, F1, and AUROC in Figure 10 in the appendix, one can observe that, when just focusing on a single metric such as Accuracy when evaluating and comparing the classifier's performance, TimeMIL performs significantly better than the other models. However, when looking at the score distributions of the F1 and AUROC metrics, this seems less evident in the case of TodyNet. This visual example underscores the hypothesis test results, which state that TimeMIL is not significantly better than TodyNet for the Twitter dataset, emphasizing why a quantitative evaluation of models is strictly necessary when choosing the appropriate modeling approach. Regarding ***H1*** which stated superior predictive performance of MIL compared to supervised models, the results of the experiments provide partial support for the hypothesis that weakly-supervised MIL models, such as TimeMIL, outperform supervised time series classification methods in classifying movies into success categories. Specifically, TimeMIL demonstrated superiority over the LSTM model on the Twitter dataset, showcasing its ability to handle social media data sources such as tweets effectively. This aligns with the ground assumption of the weakly-supervised framework, as weakly-supervised approaches can leverage unlabeled data and capitalize on the diversity and noisiness of information. However, TimeMIL's performance on other datasets, such as Rotten Tomatoes and IMDb, was

consistently inferior to that of the supervised models, particularly TodyNet. TodyNet achieved significantly higher F1 scores and AUROC values for these datasets, demonstrating that supervised models may be advantageous when handling movie ratings and Box office success data as time series. Furthermore, while TimeMIL marginally outperformed TodyNet on the Twitter dataset, this difference was not statistically significant, suggesting that TimeMIL's strength on partially labeled data is less pronounced than more robust supervised architecture like TodyNet. These findings highlight the contextual strengths and limitations of weakly-supervised MIL models. While they excel in leveraging unlabeled, potentially noisy data, applying them to more other datasets and scenarios remains challenging. This suggests that supervised models, especially those optimized for specific data domains, may outperform weakly-supervised models in many real-world scenarios. Consequently, the hypothesis **(*H1*)** is not supported, as TimeMIL's advantage seems dataset-dependent rather than universal, underscoring the importance of selecting model architectures tailored to the nature of the data in time series classification tasks. Generally, comparing different deep learning architectures and the differences in the underlying deep learning assumptions for interpretability was a rather difficult task. Suitable and feasible to compute metrics were rather difficult to implement, picking up on existing works within the field of explainable AI (Nguyen & Martínez, 2020; Turbé et al., 2023). This is why a fusion of a set of concepts was applied, which considered concepts of corresponding works within the field of explainable AI and model interpretability. The interpretability results do not support ***H2***, which proposed that weakly-supervised MIL models like TimeMIL offer more interpretable insights into movie success factors than supervised models. TimeMIL failed to outperform traditional supervised models such as TodyNet and LSTM, showing higher infidelity scores and less focused attributions. Consistently, TimeMIL had worse infidelity across datasets indicating that it cannot compete with supervised techniques when it comes to time series classification for movie success prediction. These findings

indicate that supervised models, particularly TodyNet, are more effective for generating faithful and interpretable attributions in classifying movie success.

5.6. Limitations & Future Research

Although the datasets in this study provided diverse perspectives, constructing a well-defined dataset was challenging. The available data samples across rating platforms limited the final predictive performance of the models. Additionally, the models' predictive limitations can also be argued from a technical perspective. Regarding the input sequence preprocessing, the challenge of fixed sequence lengths could be addressed in the future by exploring variable lengths, incorporating adaptive mechanisms during training, and treating sequence length as a tunable hyperparameter for optimization. Moreover, the dataset-specific results highlight the importance of selecting appropriate DL architectures. Thus, future work could explore additional DL frameworks and architectures for movie success prediction using time series data. The interpretability assessment, constrained by a limited set of attribution methods evaluated through the Infidelity metric, faced several limitations. As noted by Turbé et al. (2023) certain models may appear less interpretable due to poor alignment with the chosen attribution techniques rather than their inherent properties. Aligning with the results presented in their paper, Shapley and SHAP appeared to be less applicable in the chosen comparison setup, as both techniques are less scalable than the attribution methods used. Further, only the Infidelity metric was evaluated neglecting other methods. However, it highly depends on the selected perturbation scheme, which raises questions about generalizability across different methods and models (Turbé et al., 2023). Consequently, important dimensions of interpretability, such as sensitivity, simplicity, and stability, remain unexplored in this context (Nguyen & Martínez, 2020; H. Zhang et al., 2020). Future work should test a broader range of attribution techniques, adopt more robust perturbation frameworks, and incorporate additional interpretability metrics to gain a more holistic view of model interpretability.

6. Conclusion & Implications for the Movie Industry

The movie industry presents a compelling landscape for the application of advanced analytics and machine learning frameworks due to its reliance on diverse data types and sources. This study explores the use of both structured and unstructured data, demonstrating how these distinct forms of information can be harnessed to address various challenges and opportunities within the domain. Traditional machine learning methods were employed on structured tabular data, such as box office revenues**,** to build predictive models and identify critical factors influencing movie success. For unstructured data, natural language processing (NLP) techniques were applied to textual information such as movie reviews to predict ratings, uncover sentiment patterns, and derive deeper insights into audience preferences. Additionally, the analysis of predictive models for multivariate time series data, encompassing features for movies ratings, review sentiment, box office revenues, provided a dynamic perspective how data around audience behavior and market fluctuations can be applied in novel frameworks. By capturing how these features evolve and interact across temporal dimensions, the study highlights the potential of applying regression analysis, NLP and fusion of sequential models with interpretability techniques to understanding trends, forecasting demand, and adapting strategies to optimize outcomes in the highly competitive movie industry. This integrative approach underscores the importance of leveraging multiple data modalities, each contributing unique insights that collectively enhance decision-making and operational efficiency. Building on the diverse data types, this study implemented a range of machine learning frameworks to extract actionable insights. Supervised learning methods, including regression analysis and classification models, were applied to structured tabular data to predict outcomes such as box office success or genre-specific performance. For text data, deep learning techniques leveraging pretrained transformer models, such as BERT, were employed. These models enabled the extraction

of contextual insights and enhanced the prediction of ratings, sentiment analysis, and audience preferences by capturing the nuanced semantics of textual content. In contrast, weakly-supervised learning frameworks, such as Multiple Instance Learning (MIL), were compared with supervised models for the case of multivariate post-release time series data, where complete labeling was challenging. Critic ratings emerged as a more significant driver of box office performance than user ratings, particularly for smaller studios and off-season releases. Positive critical reviews served as vital quality signals, especially for films with limited brand recognition. By contrast, user ratings showed limited predictive power, likely due to their variability and susceptibility to biases. Seasonal trends amplified the importance of critic ratings, with off-season films relying more heavily on professional reviews to attract audiences. These findings highlight the enduring influence of expert evaluations and the importance of timing in maximizing movie success. For studios, strategically fostering relationships with critics and aligning release schedules with periods of lower competition can amplify box office returns, particularly for niche or independent productions. The impact of releasing movies during holiday periods is highly audience dependent. For non-adult movies, holiday releases significantly boost box-office revenues, aligning with increased audience availability and family viewing habits. Conversely, adult movies experience insignificant marginal or negative effects during holidays, likely due to competition with family-oriented films. Including holidays as a predictive factor shows limited overall significance when considering all movie types collectively. These findings suggest that strategic release timing tailored to audience demographics is essential. For the film industry, leveraging holiday periods for family-friendly content while reserving less competitive windows for adult-oriented films could optimize revenues, refine marketing strategies, and enhance overall return on investment. Sequel modeling revealed that audience engagement, measured by review volume, was the most influential factor in predicting both absolute and relative success. Advanced machine learning models, such

as CatBoost, outperformed traditional approaches, highlighting the importance of capturing non-linear relationships and feature interactions. Metrics like runtime and sentiment played secondary roles, with engagement metrics proving far more predictive of financial outcomes. For the movie industry, these results emphasize the need to maintain sustained audience interaction through marketing campaigns and social media presence, particularly for franchise films. Studios can use these predictive models to assess the financial viability of sequels, optimize resource allocation, and design strategies that prioritize audience retention. Transfer learning transformer-based models, particularly BERT-based frameworks, proved effective in extracting nuanced insights from textual data, achieving high performance in sentiment analysis and rating prediction. Notably, classification models benefited from strong sentiment polarity, while regression models demonstrated resilience across varied sentiment distributions. These findings suggest that advanced NLP techniques can enable movie studios to better understand audience feedback and preferences, facilitating more precise targeting and informed creative decisions. For the industry, this underscores the potential of leveraging automated sentiment analysis tools to refine marketing strategies, improve audience engagement, and enhance content development based on data-driven insights. When applied to multivariate time series data, Multiple Instance Learning (MIL) did not perform as effectively as supervised models like TodyNet in terms of predictive accuracy. Although MIL sowcased better performance particularly in unstructured data applications like Twitter, it did not outperform supervised methods signaling that movie data might not be the perfect fit. These results suggest that while interpretability-focused techniques like MIL can be applied to data involved in decision-making processes for movies, supervised models remain the preferred choice both in terms of predictive accuracy and interpretability. However, for the movie industry, integrating such techniques into analytics pipelines in scenarios where labels are scarce could improve the understanding of temporal dynamics, such as post-release audience behavior or the

impact of streaming trends, enabling studios to make more data-informed distribution and marketing decisions. This comprehensive machine learning framework covering a wide range of datatypes and applications in the movie industry proves solid potential. This could be further enhanced by adopting a multimodal approach by integrating diverse data sources to develop more comprehensive and robust analytical frameworks that go beyond parallel model application but leverage various strengths across a sophisticated system (Madongo & Zhongjun, 2023; Miah et al., 2024). By combining structured data, such as box office performance, audience demographics, and release schedules, with unstructured data from movie reviews, sentiment, and visual or audio content, researchers can unlock synergies across data modalities that provide a more nuanced understanding of audience behavior and market dynamics. Further leveraging additional data types like image and video and integrating them into NLP techniques could enable more precise insights into the impact of soundscapes on viewer experiences. Such multimodal frameworks would allow for a broader analytical bandwidth, enhancing predictive performance and interpretability in areas (Mühling et al., 2017; Tahmasebi et al., 2021; Zhou et al., 2019). To advance the application of analytics in a multimodal fashion accounting for different data representations, academic research can contribute to more sophisticated decision-making processes and innovative strategies for addressing the dynamic challenges of the movie industry. In conclusion, by applying advanced machine learning techniques in the movie industry, this study presents a transformative opportunity for the domain to address evolving challenges and capitalize on emerging trends. Through combination of structured and unstructured data with predictive frameworks, industry stakeholders can develop strategies that enhance production efficiency, optimize release schedules, and better cater to audience preferences. This study highlights the critical role of machine learning in enabling studios to remain competitive in an increasingly data-driven landscape, urging the adoption of advanced analytics to navigate the complexities of modern cinema.

7. Limitations & Future Research

This study provides valuable insights into the factors influencing movie performance, yet several limitations need to be addressed to fully contextualize the findings and guide future research. These limitations arise from challenges in data acquisition, feature availability, methodological approaches, and computational constraints, which collectively influence the scope and depth of the analyses.

The studies either used a single or a combined selection of data sources, such as Rotten Tomatoes, IMDb, Twitter or BoxOfficeMojo, to create their input data, picking up on existing work (Subramaniyaswamy, Vaibhav, Prasad, & Logesh, 2017). As showcased by Y. Liao et al. (2022), one must point out that other sources exist to retrieve movie-related data that could potentially be used to conduct research, such as data from the Douban movie platform, search engine data such as Baidu, or data from other Microblogs. However, the scope of data collection was constrained by reliance on open-source platforms, which limited the comprehensiveness of the datasets. Specifically, box office data was predominantly focused on North American markets, excluding the broader global landscape and its associated cultural and regional variations. This narrow focus limits the generalizability of findings to diverse geographical contexts. Future research should aim to expand data acquisition efforts to incorporate global box office data and explore differences in audience behaviors across distinct geographical and cultural landscapes, providing a more holistic understanding of movie performance drivers.

Several important features that are likely to influence movie success were unavailable, notably marketing and advertising budgets. These variables are critical in shaping audience awareness and engagement and could significantly enhance predictive models. Additionally, the datasets exhibited heavy skews in certain features, such as genres, which may overrepresent popular categories while

underrepresenting niche ones. Ratings data, another key feature, is inherently subjective and susceptible to biases from both users and critics, potentially affecting the accuracy of performance predictions. Addressing these limitations through more diverse and balanced data sources would greatly improve future analyses. While the machine learning models employed in this study demonstrated strong predictive capabilities, they were inherently limited in their ability to establish causal relationships due to inherit irregularity within the movies data. Identifying correlations between variables is valuable for predictive analytics, but without causal inference, it remains difficult to determine which factors directly drive movie success. A notable limitation across the studies was the lack of detailed audience segmentation. While some analyses differentiated between adult and non-adult movies, there was limited exploration of granular divisions based on demographics, cultural preferences, or regional behaviors. This lack of segmentation restricts the ability to identify specific audience behaviors and preferences, which could be critical for tailoring marketing strategies and release decisions. Future research should prioritize deeper audience segmentation to uncover nuanced patterns influencing movie performance. Limited computational resources affected the ability to explore advanced techniques, such as dynamic modeling, causal inference methods, or the use of larger and more sophisticated transformer architectures. These constraints also restricted the granularity of fine-tuning and optimization processes, potentially limiting the models' predictive accuracy. Addressing these computational challenges in future work would enable more robust analyses, such as the inclusion of broader datasets, advanced model architectures, and more nuanced hyperparameter optimization. This research paper provides a detailed exploration of various factors influencing movie performance, with each study contributing particular insights into predictive analytics around movies, consumer behavior, and strategic decision-making processes within the movie industry.

8. Bibliography

Ahmad, J., Duraisamy, P., Yousef, A., & Buckles, B. (2017). Movie success prediction using data mining. *8th International Conference on Computing, Communication and Networking Technologies*, 1–4. IEEE. https://doi.org/10.1109/ICCCNT.2017.8204173

Ahmed, B. H., & Ghabayen, A. S. (2022). Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, *13*(7), 3423–3432. https://doi.org/10.1007/s12652-020-01807-4

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework* (pp. 1–10). pp. 1–10. Retrieved from http://arxiv.org/abs/1907.10902

Aljrees, T., Umer, M., Saidani, O., Almuqren, L., Ishaq, A., Alsubai, S., … Ashraf, I. (2024). Contradiction in text review and apps rating: Prediction using textual features and transfer learning. *PeerJ Computer Science*, *10*(7), 1–25. https://doi.org/10.7717/peerj-cs.1722

Al-Kababji, A., Bensaali, F., & Dakua, S. P. (2022). *Scheduling techniques for liver segmentation: ReduceLROnPlateau vs OneCycleLR* (pp. 1–8). pp. 1–8. Retrieved from http://arxiv.org/abs/2202.06373

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509. https://doi.org/10.1287/mnsc.1110.1370

Assis, A., Dantas, J., & Andrade, E. (2024). The performance-interpretability trade-off: A comparative study of machine learning models. *Journal of Reliable Intelligent Environments*, *11*(1), 1–15. https://doi.org/10.1007/s40860-024-00240-0

Bagnall, A., Dau, H. A., Lines, J., Flynn, M., Large, J., Bostrom, A., … Keogh, E. (2018). *The UEA multivariate time series classification archive*. Retrieved from http://arxiv.org/abs/1811.00075

Bakdi, A., Kristensen, N. B., & Stakkeland, M. (2022). Multiple instance learning with random forest for event logs analysis and predictive maintenance in ship electric Propulsion System. *IEEE Transactions on Industrial Informatics*, *18*(11), 7718–7728. https://doi.org/10.1109/TII.2022.3144177

Banbhrani, S. K., Xu, B., Soomro, P. D., Jain, D. K., & Lin, H. (2022). TDO-Spider Taylor ChOA: An optimized deep-learning-based sentiment classification and review rating prediction. *Applied Sciences*, *12*(20), 1–26. https://doi.org/10.3390/app122010292

Banik, R. (2017). The Movies Dataset IMDb. Retrieved December 13, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download&select=links.csv

Basuroy, S. D. K. K. T. D. (2006). An empirical investigation of signaling in the motion picture industry. *Journal of Marketing Research*, *43*(2), 287–295. https://doi.org/https://doi.org/10.1509/jmkr.43.2.287

Belvaux, B., & Mencarelli, R. (2021). Prevision model and empirical test of box office results for sequels. *Journal of Business Research*, *130*, 38–48. https://doi.org/10.1016/j.jbusres.2021.03.008

Berger, J., Kim, Y. D., & Meyer, R. (2021). What makes content engaging? How emotional dynamics shape success. *Journal of Consumer Research*, *48*(2), 235–250. https://doi.org/10.1093/jcr/ucab010

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). Is attention explanation? An introduction to the debate. *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics, *1*, 3889–3900. Long Papers. https://doi.org/10.18653/v1/2022.acl-long.269

Bing, L., & Wang, W. (2017). Sparse representation based multi-instance learning for breast ultrasound image classification. *Computational and Mathematical Methods in Medicine*, *2017*(1), 1–10. https://doi.org/10.1155/2017/7894705

Bureau of Labor. (2023). Occupational employment and wage statistics. Retrieved November 12, 2024, from https://www.bls.gov/oes/2023/may/naics4_512100.htm

Burton, A. L. (2021). OLS (linear) regression. In *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (pp. 1–11). Wiley. https://doi.org/DOI:10.1002/9781119111931

Carbonneau, M. A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, *77*(2), 329–353. https://doi.org/10.1016/j.patcog.2017.10.009

Chambua, J., & Niu, Z. (2021). Review text based rating prediction approaches: Preference knowledge learning, representation and utilization. *Artificial Intelligence Review*, *54*(2), 1171–1200. https://doi.org/10.1007/s10462-020-09873-y

Chen, X., Qiu, P., Zhu, W., Li, H., Wang, H., Sotiras, A., … Razi, A. (2024). *TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning* (pp. 1–17). pp. 1–17. Retrieved from http://arxiv.org/abs/2405.03140

Chung, C., Niu, S. C., & Sriskandarajah, C. (2012). A sales forecast model for short-life-cycle products: New releases at blockbuster. Production and Operations Management, 21(5), 851-873. https://doi.org/10.1111/j.1937-5956.2012.01326.x

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Mehmood, F., & Ali, I. (2024). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. *Multimedia Tools and Applications*, *83*(24), 64315–64339. https://doi.org/10.1007/s11042-024-18156-5

Darwish, A., Hassanien, A. E., & Das, S. (2020). A survey of swarm and evolutionary computing approaches for deep learning. *Artificial Intelligence Review*, *53*(3), 1767–1812. https://doi.org/10.1007/s10462-019-09719-2

D'astous, A., Montreal, H., Touil, N., & Tunisia, P. (1999). Consumer evaluations of movies on the basis of critics' judgments. *Psychology & Marketing*, *16*(8), 677–694. https://doi.org/https://doi.org/10.1002/(SICI)1520-6793(199912)16:8<677::AID-MAR4>3.0.CO;2-T

del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2017). An optimal transportation approach for assessing almost stochastic order. In *Studies in Systems, Decision and Control* (Vol. 142, pp. 33–44). Springer International. https://doi.org/https://doi.org/10.1007/978-3-319-73848-2_3

Dellarocas, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45. https://doi.org/10.1002/dir.20087

Delre, S. A., & Luffarelli, J. (2023). Consumer reviews and product life cycle: On the temporal dynamics of electronic word of mouth on movie box office. *Journal of Business Research*, *156*, 1–26. https://doi.org/10.1016/j.jbusres.2022.113329

Deng, Y., & Yiu, S. M. (2022). Deep multiple instance learning for forecasting stock trends using financial NewsDeep multiple instance learning for forecasting stock trends using financial news. *Conference: 8th International Conference on Artificial Intelligence*, 95–111. Academy and Industry Research Collaboration Center (AIRCC). https://doi.org/10.5121/csit.2022.121008

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved from https://github.com/tensorflow/tensor2tensor

Divakaran, P. K. P., & Nørskov, S. (2016). Are online communities on par with experts in the evaluation of new movies? Evidence from the fandango community. *Information Technology and People*, *29*(1), 120–145. https://doi.org/10.1108/ITP-02-2014-0042

Divakaran, P. K. P., Palmer, A., Sendergaard, H. A., & Matkovskyy, R. (2017). Pre-launch prediction of market performance for short lifecycle products using online community data. *Journal of Interactive Marketing*, *38*(1), 12–28. https://doi.org/https://doi.org/10.1016/j.intmar.2016.10.004

Dooms, S. (2021). MovieTweetings. Retrieved December 13, 2024, from https://github.com/sidooms/MovieTweetings

Dror, R., Shlomov, S., & Reichart, R. (2019). Deep dominance - how to properly compare deep neural models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2773–2785. Association for Computational Linguistics. https://doi.org/https://doi.org/10.18653/v1/p19-1266

Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. *Journal of Retailing*, *84*(2), 233–242. https://doi.org/10.1016/j.jretai.2008.04.005

Early, J., Cheung, G. K., Cutajar, K., Xie, H., Kandola, J., & Twomey, N. (2024). *Inherently interpretable time series classification via multiple instance learning*. Retrieved from http://arxiv.org/abs/2311.10049

Einav, L. (2007). Seasonality in the U.S. motion picture industry. *RAND Journal of Economics*, *38*(1), 127–145. https://doi.org/10.1111/j.1756-2171.2007.tb00048.x

Flanagin, A. J., & Metzger, M. J. (2013). Trusting expert versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, *29*(4), 1626–1634. https://doi.org/10.1016/j.chb.2013.02.001

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). SAGE Publications.

Galal, O., Abdel-Gawad, A. H., & Farouk, M. (2024). Rethinking of BERT sentence embedding for text classification. *Neural Computing and Applications*, *36*, 20245–20258. https://doi.org/10.1007/s00521-024-10212-3

García, S., Herrera, F., & Es, H. U. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, *9*(12), 2677–2694. https://doi.org/https://jmlr.org/papers/v9/garcia08a.html

Ghanbari, R., & Borna, K. (2021). Multivariate Time-Series Prediction Using LSTM Neural Networks. *26th International Computer Conference, Computer Society of Iran, CSICC 2021*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CSICC52343.2021.9420543

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: An overview*. Retrieved from http://arxiv.org/abs/2008.05756

Haben, S., Voss, M., & Holderbaum, W. (2023). Time series forecasting: Core concepts and definitions. In *Core concepts and methods in load forecasting* (pp. 55–66). Springer International. https://doi.org/https://doi.org/10.1007/978-3-031-27852-5_5

He, H., Wang, H., Ma, H., Liu, X., Jia, Y., & Gong, G. (2020). Research on short-term power load forecasting based on Bi-GRU. *Journal of Physics: Conference Series*, *1639*(1), 1–7. IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1639/1/012017

Hennig-Thurau, T., Houston, M. B., & Heitjans, T. (2009). Conceptualizing and measuring the monetary value of brand extensions: The case of motion Pictures. *Journal of Marketing*, *73*(6), 167–183. https://doi.org/10.1509/jmkg.73.6.167

Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, *99*(1), 4806–4813. https://doi.org/10.1109/ACCESS.2019.2962617

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Jain, S., & Wallace, B. C. (2019). *Attention is not Explanation*. https://doi.org/https://doi.org/10.18653/v1/D19-1002

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*, 685–695. https://doi.org/10.1007/s12525-021-00475-2/Published

Karniouchina, E. V., Carson, S. J., Theokary, C., Rice, L., & Reilly, S. (2023). Women and minority film directors in hollywood: Performance implications of product development and distribution biases. *Journal of Marketing Research*, *60*(1), 25–51. https://doi.org/10.1177/00222437221100217

Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, *4*(2), 1–19. https://doi.org/10.1186/s40294-016-0016-9

Khan, Z. Y., Niu, Z., Sandiwarno, S., & Prince, R. (2021). Deep learning techniques for rating prediction: A survey of the state-of-the-art. *Artificial Intelligence Review*, *54*(1), 95–135. https://doi.org/10.1007/s10462-020-09892-9

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., … Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for PyTorch* (pp. 1–11). pp. 1–11. https://doi.org/10.48550/arXiv.2009.07896

Kolpaczki, P., Bengs, V., Muschalik, M., & Hüllermeier, E. (2024). *Approximating the Shapley Value without Marginal Contributions* (pp. 1–37). pp. 1–37. https://doi.org/https://doi.org/10.48550/arXiv.2009.07896

Kraft, A., & Rao, R. S. (2024). Signaling quality via demand lockout. *Quantitative Marketing and Economics*. https://doi.org/10.1007/s11129-024-09288-x

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, *33*(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Lee, J. H., Jung, S. H., & Park, J. H. (2017). The role of entropy of review text sentiments on online WOM and movie box office sales. *Electronic Commerce Research and Applications*, *22*(4), 42–52. https://doi.org/10.1016/j.elerap.2017.03.001

Lehrer, S. F., & Xie, T. (2021). *The bigger picture: Combining econometrics with analytics improve forecasts of movie success* (No. 24755). Retrieved from http://www.nber.org/papers/w24755

Leone, S. (2020). Rotten Tomatoes movies and critic reviews dataset. Retrieved December 13, 2024, from https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/

Liao, L., & Huang, T. (2021). The effect of different social media marketing channels and events on movie box office: An elaboration likelihood model perspective. *Information and Management*, *58*(7), 1–16. https://doi.org/10.1016/j.im.2021.103481

Liao, Y., Peng, Y., Shi, S., Shi, V., & Yu, X. (2022). Early box office prediction in china's film market based on a stacking fusion model. *Annals of Operations Research*, *308*(4), 321–338. https://doi.org/10.1007/s10479-020-03804-4

Liu, H., Yang, D., Liu, X., Chen, X., Liang, Z., Wang, H., … Gu, J. (2024). TodyNet: Temporal dynamic graph neural network for multivariate time series classification. *Information Sciences*, *677*, 1–16. https://doi.org/10.1016/j.ins.2024.120914

Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization*. 1–19. https://doi.org/https://doi.org/10.48550/arXiv.1711.05101

Loupos, P., Peng, Y., Li, S., & Hao, H. (2023). What reviews foretell about opening weekend box office revenue: the harbinger of failure effect in the movie industry. *Marketing Letters*, *34*(3), 513–534. https://doi.org/10.1007/s11002-023-09665-8

Madongo, C. T., & Zhongjun, T. (2023). A movie box office revenue prediction model based on deep multimodal features. *Multimedia Tools and Applications*, *82*(21), 31981–32009. https://doi.org/10.1007/s11042-023-14456-4

Miah, M. S. U., Kabir, M. M., Sarwar, T. Bin, Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, *14*(1), 1–18. https://doi.org/10.1038/s41598-024-60210-7

Middlehurst, M., Ismail-Fawaz, A., Guillaume, A., Holder, C., Guijo-Rubio, D., Bulatova, G., … Bagnall, A. (2024). Aeon: A python toolkit for learning from time series. *Journal of Machine Learning Research*, *25*, 1–10. Retrieved from http://jmlr.org/papers/v25/23-1444.html.

Mohamadi, P. (2023). *Unofficial Python API for Box Office Mojo*. Retrieved from https://github.com/Stink-Po/boxoffice_api

Mohamadi, P. (2024). Box Office Mojo API. Retrieved December 13, 2024, from https://github.com/Stink-Po/boxoffice_api

Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, *74*(1), 108–121. https://doi.org/10.1509/jmkg.74.1.108

Mosin, V., Samenko, I., Kozlovskii, B., Tikhonov, A., & Yamshchikov, I. P. (2023). Fine-tuning transformers: Vocabulary transfer. *Artificial Intelligence*, *317*, 1–11. https://doi.org/10.1016/j.artint.2023.103860

Mühling, M., Korfhage, N., Müller, E., Otto, C., Springstein, M., Langelage, T., … Freisleben, B. (2017). Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, *76*(21), 22169–22194. https://doi.org/10.1007/s11042-017-4962-9

Namdari, A., & Li, Z. (Steven). (2019). A review of entropy measures for uncertainty quantification of stochastic processes. *Advances in Mechanical Engineering*, *11*(6), 1–14. https://doi.org/10.1177/1687814019857350

Navarathna, R., Carr, P., Lucey, P., & Matthews, I. (2019). Estimating audience engagement to predict movie ratings. *IEEE Transactions on Affective Computing*, *10*(1), 48–59. https://doi.org/10.1109/TAFFC.2017.2723011

Nguyen, A., & Martínez, M. R. (2020). *On quantitative aspects of model interpretability* (pp. 1–14). pp. 1–14. https://doi.org/https://doi.org/10.48550/arXiv.2007.07584

Pang, J., Liu, A. X., & Golder, P. N. (2022). Critics' conformity to consumers in movie evaluation. *Journal of the Academy of Marketing Science*, *50*(4), 864–887. https://doi.org/10.1007/s11747-021-00816-9

Park, S. K., Song, T., & Sela, A. (2023). The effect of subjectivity and objectivity in online reviews: A convolutional neural network approach. *Journal of Consumer Psychology*, *33*(4), 701–713. https://doi.org/10.1002/jcpy.1382

Perez, I., Skalski, P., Barns-Graham, A., Wong, J., & Sutton, D. (2021). *Attribution of predictive uncertainties in classification models* (pp. 1–16). pp. 1–16. https://doi.org/https://doi.org/10.1109/ICCV.2019.00505

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. In *Scientific Reports* (Vol. 14). Nature Research. https://doi.org/10.1038/s41598-024-56706-x

Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2021). BERT, XLNet or RoBERTa: The best transfer learning model to detect clickbaits. *IEEE Access*, *9*, 154704–154716. https://doi.org/10.1109/ACCESS.2021.3128742

Ren, Z., Wang, S., & Zhang, Y. (2023). Weakly supervised machine learning. *CAAI Transactions on Intelligence Technology*, *8*(3), 549–580. https://doi.org/10.1049/cit2.12216

Rhee, T., & Zulkerine, F. H. (2016). Predicting movie box office profitability: A neural network approach. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 665–670. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICMLA.2016.138

Ryoo, J. H., Wang, X., & Lu, S. (2021). Do spoilers really spoil? Using topic modeling to measure the effect of spoiler reviews on box office revenue. *Journal of Marketing*, *85*(2), 70–88. https://doi.org/10.1177/0022242920937703

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., … Coyle, D. (2024). *Computing power and the Governance of artificial intelligence* (pp. 1–104). pp. 1–104. https://doi.org/https://doi.org/10.48550/arXiv.2402.08797

Schlegel, U., & Keim, D. A. (2021). *Time series model attribution visualizations as explanations* (pp. 1–5). pp. 1–5. https://doi.org/https://doi.org/10.48550/arXiv.2109.12935

Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845. https://doi.org/10.3758/s13428-023-02072-x

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). *Learning important features through propagating activation Differences* (pp. 1–9). pp. 1–9. https://doi.org/https://doi.org/10.48550/arXiv.1704.02685

Simonton, D. K. (2009). Cinematic success criteria and their predictors: The art and business of the film industry. *Psychology and Marketing*, *26*(5), 400–420. https://doi.org/10.1002/mar.20280

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency Maps* (pp. 1–8). pp. 1–8. https://doi.org/https://doi.org/10.48550/arXiv.1312.6034

Somlo, B., Rajaram, K., & Ahmadi, R. (2011). Distribution planning to optimize profits in the motion picture industry. *Production and Operations Management*, *20*(4), 618–636. https://doi.org/10.1111/j.1937-5956.2010.01166.x

Song, T., Huang, J., Tan, Y., & Yu, Y. (2019). Using user- and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Information Systems Research*, *30*(1), 191–203. https://doi.org/10.1287/isre.2018.0797

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 182–186. Palladam: IEEE. https://doi.org/10.1109/ISS1.2017.8389394

Subramaniyaswamy, V., Viginesh, V. M., Vishnu, P. R., & Logesh, R. (2017). Predicting movie box office success using multiple regression and svm. *Proceedings of the International Conference*

on Intelligent Sustainable Systems, 526. Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ISS1.2017.8389394

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks* (pp. 1–11). pp. 1–11. https://doi.org/https://doi.org/10.48550/arXiv.1703.01365

Tahmasebi, H., Ravanmehr, R., & Mohamadrezaei, R. (2021). Social movie recommender system based on deep autoencoder network using Twitter data. *Neural Computing and Applications*, *33*(5), 1607–1623. https://doi.org/10.1007/s00521-020-05085-1

The Numbers. (2024). All time worldwide sequel box office. Retrieved December 13, 2024, from https://www.the-numbers.com/box-office-records/worldwide/all-movies/cumulative/sequel

Tibo, A., Jaeger, M., Frasconi, P., & Wrobel, S. (2020). Learning and interpreting multi-multi-instance learning networks. *Journal of Machine Learning Research*, *21*, 1–60. https://doi.org/https://www.jmlr.org/papers/v21/18-811.html

TMDB. (2023). TMDB API. Retrieved December 13, 2024, from https://developer.themoviedb.org/reference/intro/getting-started

TMDB API (Vol. 3). (2023). TiVo Platform Technologies LLC. Retrieved from https://developer.themoviedb.org/docs/getting-started

Tsao, W. C. (2014). Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers. *Electronic Commerce Research*, *14*(4), 559–583. https://doi.org/10.1007/s10660-014-9160-5

Turbé, H., Bjelogrlic, M., Lovis, C., & Mengaldo, G. (2023). Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, *5*(3), 250–260. https://doi.org/10.1038/s42256-023-00620-w

Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). *Easy and meaningful statistical significance testing in the age of neural networks* (pp. 1–20). pp. 1–20. Retrieved from http://arxiv.org/abs/2204.06815

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., … Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems*, 1–11.

Wadawadagi, R., & Pagi, V. (2020). Sentiment analysis with deep neural networks: Comparative study and performance assessment. *Artificial Intelligence Review*, *53*(8), 6155–6195. https://doi.org/10.1007/s10462-020-09845-2

Wang, H., & Guo, K. (2017). The impact of online reviews on exhibitor behaviour: Evidence from movie industry. *Enterprise Information Systems*, *11*(10), 1518–1534. https://doi.org/10.1080/17517575.2016.1233458

Watson, F., & Wu, Y. (2022). The Impact of online reviews on the information flows and outcomes of marketing systems. *Journal of Macromarketing*, *42*(1), 146–164. https://doi.org/10.1177/02761467211042552

Yang, G., Xu, Y., & Tu, L. (2023). An intelligent box office predictor based on aspect-level sentiment analysis of movie review. *Wireless Networks*, *29*(7), 3039–3049. https://doi.org/10.1007/s11276-023-03378-6

Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). *On the (in)fidelity and sensitivity for explanations*. https://doi.org/https://doi.org/10.48550/arXiv.1901.09392

Zhang, C., Tian, Y. X., & Fan, Z. P. (2022). Forecasting the box offices of movies coming soon using social media analysis: A method based on improved bass models. *Expert Systems with Applications*, *191*, 1–16. https://doi.org/10.1016/j.eswa.2021.116241

Zhang, H., Yuan, X., & Song, T. H. (2020). Examining the role of the marketing activity and eWOM in the movie diffusion: The decomposition perspective. *Electronic Commerce Research*, *20*(3), 589–608. https://doi.org/10.1007/s10660-020-09423-2

Zhou, Y., Zhang, L., & Yi, Z. (2019). Predicting movie box-office revenues using deep neural networks. *Neural Computing and Applications*, *31*(6), 1855–1865. https://doi.org/10.1007/s00521-017-3162-x

9. Appendix

9.1. Temporal Modeling for Movie Success Prediction: A Comparative Study on the Applicability of different Deep Learning Models in Entertainment Analytics

Equation 1: Movies Return-on-Invest

$$RoI_i = \frac{Revenue_i - Budget_i}{Budget_i}$$

Where $i \in M$: $M\ is\ the\ set\ of\ movies\{i, \dots, n\}$.

$Revenue_i$: Total revenue generated by the movie i over its lifetime.

$Budget_i$: The production cost or budget for movie i.

Table 5: Summary Statistic for Time Series Daatsets before and after Preprocessing.

Dataset *Source*	**Raw TS Dimensionality** *(num rows/num cols)*	**Aggregated + Filtered TS Dimensionality** *(num rows/num cols)*	**Number Movies** *#*	**Seq. Length** *Median*	**Target Distribution** *(hit; avg; flop)*
Rotten	(1009126, 7)	(16687, 15)	450	34	(0.47; 0.29; 0.24)
IMDB	(26024289, 6)	(96971, 12)	1365	57	(0.44; 0.35; 0.21)
Twitter	(921398, 6)	(26848, 11)	393	59	(0.67; 0.28; 0.05)
BoxOffice	(309561, 4)	(309561, 4)	5111	60	-

Table 6: Features included in each of the three datasets.

Rotten Tomatoes	**IMDB**	**Twitter**
Revenue	*Revenue*	*Revenue*
Budget	*Budget*	*Budget*
MovieID	*MovieID*	*MovieID*
Date	*Date*	*Date*
rating_score_encoded_mean	*rating_score_encoded_mean*	*rating_score_encoded_mean*
rating_score_encoded_var	*rating_score_encoded_var*	*rating_score_encoded_min*
rating_score_encoded_min	*rating_score_encoded_min*	*rating_score_encoded_max*
rating_score_encoded_max	*rating_score_encoded_max*	*review_word_count_mean*
review_word_count_mean	*daily_number_of_ratings*	*daily_number_users*
review_sentiment_polarity_mean	*daily_box_office*	*daily_box_office*
review_sentiment_polarity_min	*daily_theaters_avg_gross*	*daily_theaters_avg_gross*
review_sentiment_polarity_max	*daily_num_theaters*	*daily_num_theaters*
review_sentiment_polarity_var		
daily_box_office		
daily_theaters_avg_gross		
daily_num_theaters		

Table 7: Mean AUC scores across models and Datasets

Datasets/Models	*TimeMIL*	*TodyNet*	*LSTM*	*Rank* TimeMIL
Rotten Tomatoes	0.658596	**0.749291**	0.698412	3
IMDB	0.651499	**0.763592**	0.742273	3
Twitter	**0.750273**	0.750155	0.729751	1

Table 8: Mean Precision scores across models and Datasets.

Datasets/Models	*TimeMIL*	*TodyNet*	*LSTM*	*Rank* (MIL/Tody/LSTM)
Rotten Tomatoes	0.456646	0.547320	0.508332	3
IMDB	0.476054	0.585437	0.551034	3
Twitter	0.511275	0.500420	0.457373	1

Table 9: Mean Recall scores across models and Datasets.

Datasets/Models	*TimeMIL*	*TodyNet*	*LSTM*	*Rank* (MIL/Tody/LSTM)
Rotten Tomatoes	0.469753	0.554691	0.510556	3
IMDB	0.462770	0.569559	0.557993	3
Twitter	0.506270	0.534921	0.484365	3

Table 10: Mean Accuracy scores across models and Datasets.

Datasets/Models	*TimeMIL*	*TodyNet*	*LSTM*	*Rank* (MIL/Tody/LSTM)
Rotten Tomatoes	0.479412	0.524510	0.504902	3
IMDB	0.467317	0.559675	0.551545	3
Twitter	0.650847	0.566667	0.509040	1

Table 11: Mean Loss scores across models and Datasets.

Datasets/Models	*TimeMIL*	*TodyNet*	*LSTM*	*Rank* (MIL/Tody/LSTM)
Rotten Tomatoes	1.039801	0.843371	1.011542	3
IMDB	2.854342	0.896940	0.915404	3
Twitter	0.612022	0.461628	0.567411	3

Table 12: Shapiro Wilk Test for pairwise differences distributions of F1 Score and AUROC

MIL-Model	*Metric*	*Dataset*	*Test statistic*	*p-value*	*Conclusion*
TimeMIL vs. LSTM	*F1*	*Rotten*	*0.9648*	*0.4081*	*Normal Distribution*
		IMDB	*0.9608*	*0.3245*	*Normal Distribution*
		Twitter	*0.981*	*0.8519*	*Normal Distribution*
	AUROC	*Rotten*	*0.9843*	*0.9244*	*Normal Distribution*
		IMDB	*0.9654*	*0.4228*	*Normal Distribution*
		Twitter	*0.9678*	*0.4799*	*Normal Distribution*
TimeMIL vs. TodyNet	*F1*	*Rotten*	*0.9607*	*0.3222*	*Normal Distribution*
		IMDB	*0.9821*	*0.8772*	*Normal Distribution*
		Twitter	*0.9658*	*0.4318*	*Normal Distribution*
	AUROC	*Rotten*	*0.9746*	*0.6696*	*Normal Distribution*
		IMDB	*0.9775*	*0.7546*	*Normal Distribution*
		Twitter	*0.983*	*0.8979*	*Normal Distribution*

Table 13: Mean attribution entropy for each model-dataset and attribution method.

Models	***TimeMIL***		***TodyNet***		***LSTM***	
Datasets\Entropy	***IntGrad***	***DeepLift***	***IntGrad***	***DeepLift***	***IntGrad***	***DeepLift***
Rotten Tomatoes	2.344884	2.334574	2.207041	2.213954	2.118640	2.132280
IMDB	2.020013	1.920931	1.885764	1.872083	1.737910	1.725997
Twitter	1.848057	1.841669	1.693297	1.662531	1.677392	1.686569
Cross-data-mean	2.07098	2.032391	1.92870	1.916189	1.84464	1.84828

Table 14: Results of Shapiro-Wilk test for the infidelity metric on IntGrad and Deeplift

Dataset	**Infidelity**	**Compared Models**	**Shapiro_stat**	**Shapiro_p-value**	**Shapiro_interpretation**
rotten	IntGrad	TimeMIL vs. LSTM	0.820378672	0.006795486	Non-Normal distribution
rotten	IntGrad	TimeMIL vs. TodyNet	0.786436363	0.002475413	Non-Normal distribution
rotten	DeepLift	TimeMIL vs. LSTM	0.760036879	0.001180044	Non-Normal distribution
rotten	DeepLift	TimeMIL vs. TodyNet	0.7241101	0.000455507	Non-Normal distribution
imdb	IntGrad	TimeMIL vs. LSTM	0.825790202	0.008033388	Non-Normal distribution
imdb	IntGrad	TimeMIL vs. TodyNet	0.825815397	0.008039683	Non-Normal distribution
imdb	DeepLift	TimeMIL vs. LSTM	0.77623358	0.001850902	Non-Normal distribution
imdb	DeepLift	TimeMIL vs. TodyNet	0.776327965	0.001855838	Non-Normal distribution
twitter	IntGrad	TimeMIL vs. LSTM	0.659769637	9.52703E-05	Non-Normal distribution
twitter	IntGrad	TimeMIL vs. TodyNet	0.647835889	7.25173E-05	Non-Normal distribution
twitter	DeepLift	TimeMIL vs. LSTM	0.684341223	0.000169846	Non-Normal distribution
twitter	DeepLift	TimeMIL vs. TodyNet	0.672167277	0.000127182	Non-Normal distribution

Table 15: Results of Wilcoxon signed-rank test for the infidelity metric on IntGrad and Deeplift

Dataset	Infidelity	Compared Models	Test_used	Test_stat	Test_p-value
rotten	IntGrad	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
rotten	IntGrad	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05
rotten	DeepLift	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
rotten	DeepLift	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05
imdb	IntGrad	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
imdb	IntGrad	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05
imdb	DeepLift	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
imdb	DeepLift	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05
twitter	IntGrad	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
twitter	IntGrad	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05
twitter	DeepLift	TimeMIL vs. LSTM	Wilcoxon signed-rank test	0	6.10352E-05
twitter	DeepLift	TimeMIL vs. TodyNet	Wilcoxon signed-rank test	0	6.10352E-05

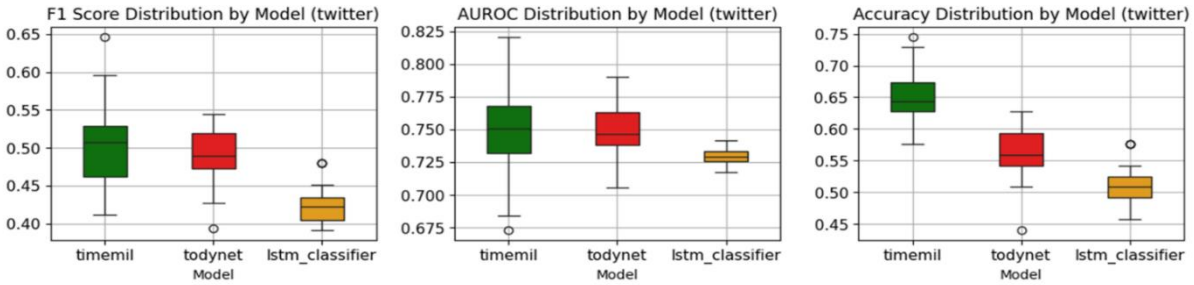


Figure 10: Distribution of Performance Metrics for each model across experiments.