iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78546

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

AI Based Popularity Prediction of TV Shows & Movies

Dr. Sajja Suneel[1], Mukkera Sarayu[2], Behnaz Mohammad[3], Pranav Anand[4]
Department of Computer Science and Engineering (Data Science) Institute of Aeronautical Engineering, Hyderabad, Telangana, India

Abstract: ***The exponential growth of digital streaming platforms has necessitated the development of precise predictive models to assess audience engagement and the anticipated popularity of television series and films prior to their release. This study introduces an AI-powered predictive system that amalgamates machine learning, natural language processing, sentiment analysis, and social media analytics to anticipate popularity metrics. The research utilises diverse data sources, including IMDB ratings, critical reviews, Twitter engagement, and box office performance. The proposed model employs feature engineering techniques such as sentiment scoring, topic modelling, and audience engagement features. Then, it uses supervised learning algorithms like Random Forest, Linear Regression, and Gradient Boosting. Experimental evaluation using accuracy, precision, recall, and F1-score demonstrates the effectiveness of the proposed framework. This paper highlights a scalable framework that can be deployed as a cloud-based API for production environments.***
Keywords: Popularity prediction, Machine learning, Sentiment analysis, Social media analytics, Movie analytics, AI forecasting.

I. INTRODUCTION

The rapid expansion of over-the-top (OTT) platforms such as Netflix, Amazon Prime, and Disney+ Hotstar has significantly transformed the global entertainment industry. Each year, a large volume of new films and television series are released, intensifying competition for audience attention. As a result, accurately forecasting content popularity has become crucial for producers, distributors, streaming platforms, and investors. Conventional prediction approaches rely heavily on subjective judgment, historical heuristics, and manually conducted surveys, which often suffer from bias, limited scalability, and inconsistent reliability. Recent advances in artificial intelligence and machine learning provide a data-driven alternative capable of processing large-scale, heterogeneous datasets. These datasets include user reviews, audience engagement metrics, trailer reactions, and online discussions across digital platforms. The objective of this work is to develop an AI- based predictive system that estimates content popularity both prior to release and during post-release phases. The proposed framework follows a structured pipeline encompassing data collection, preprocessing, feature engineering, and supervised learning model development, enabling systematic analysis of audience-driven indicators. The rapid shift toward digital media consumption has also increased the volume and velocity of user-generated content on platforms such as Twitter, YouTube, IMDb, and Reddit. These platforms play a critical role in shaping public perception well before a movie or television series is released. Modern audiences rely heavily on online sentiment, influencer opinions, teaser responses, and community discussions, making popularity a multi-dimensional concept rather than a simple measure of viewership. This complexity highlights the need for analytical models capable of capturing non-linear relationships, contextual sentiment, and interactions among diverse engagement signals. This research addresses these challenges by integrating machine learning and natural language processing techniques to analyze audience behavior at scale. By bridging traditional forecasting practices with contemporary data-intensive methodologies, the proposed approach aims to provide reliable popularity predictions that support informed decision-making within the evolving digital entertainment ecosystem.

II. RELATED WORK

A. Early Approaches to Media Popularity Prediction

Initial research in entertainment analytics primarily relied on statistical modeling. Early works used linear regression, decision trees, and time-series methods to forecast box office outcomes based on production budgets, cast popularity, and release timing. These models demonstrated essential correlations between conventional metadata and commercial success; however, their predictive efficacy was constrained by the omission of user sentiment and the dynamic nature of online activities.

©IJRASET: All Rights are Reserved

Moreover, these traditional techniques struggled to incorporate evolving viewer behaviour, limiting their ability to adapt to real-time changes in audience interests. As a result, early predictive systems often produced static or outdated forecasts that failed to capture the dynamic nature of modern media consumption.

B. Social Media Influence and User-Generated Signals

As social networks grew, researchers started looking into how user-generated content could be used to make predictions. Studies utilising Twitter buzz, hashtag trends, YouTube trailer engagements, and online discussion volumes have shown significant correlations between digital activity and real-world performance. These studies showed that social signals can often be early signs of how interested an audience is, which makes it easier to make more accurate predictions before a release. Consequently, user generated signals serve as powerful real-time indicators of audience engagement, enabling models to anticipate popularity shifts more effectively.

C. Sentiment and Review Mining Techniques

Natural language processing (NLP) became a key tool for inferring audience sentiment. Earlier systems used lexicon- based sentiment analysis tools like VADER and TextBlob. Later works used supervised classifiers and deep learning models. BERT and RoBERTa are examples of transformer- based architectures that greatly improved how well reviews could be understood in context. They also showed that sentiment is a crucial factor for predicting popularity. Research consistently demonstrates that polarity, emotional intensity, and review volume significantly affect both short- term and long-term popularity. In addition, recent advancements in contextual sentiment modelling have enabled systems to capture subtle linguistic cues such as sarcasm, comparative statements, and mixed emotions, which traditional lexicon-based tools often fail to detect. These improved interpretive capabilities help generate more precise sentiment profiles, especially for diverse genres and audience groups. Furthermore, Integrating review temporality, which captures how sentiment evolves over time, has further improved long-term popularity forecasting accuracy.

III. SYSTEM ARCHITECTURE

A. Data Ingestion Layer

The data ingestion layer acquires raw data from multiple heterogeneous sources, including IMDB, Rotten Tomatoes, Twitter, YouTube, and box office databases. APIs, web- scrapers, and database connectors are examples of automated tools that constantly gather reviews, ratings, social media posts, engagement metrics, and other data that can help predict popularity. This ensures continuous access to a diverse range of up-to-date information. To maintain consistency, the ingested data is standardized into a common format and stored in structured SQL or NoSQL databases. Basic validation and filtering are applied to remove incomplete or corrupted entries. This well-organised and consistent dataset is the basis for the preprocessing and feature engineering modules that come after it.

B. Preprocessing and Cleaning Module

This module gets the raw data ready for machine learning workflows. To make the data better, noise is removed and duplicate entries are eliminated, normalise the text, and deal with missing data. Natural language preprocessing steps like tokenisation, stop-word removal, lemmatisation, and vectorisation are used on text content like reviews and comments. To keep the scaling across the dataset consistent, numeric features like engagement metrics are normalised. To further improve consistency, outliers and irregular patterns in engagement data are smoothed or corrected. These preprocessing refinements ensure that downstream models receive high-quality inputs free from noise-induced distortions.

C. Feature Engineering and Representation Layer

This layer transforms preprocessed data and turns it into useful features. The core feature set includes sentiment scores, topic clusters, review distributions, social engagement metrics, cast popularity indicators, and metadata-derived attributes. We use TF-IDF, word embeddings, sentiment polarity scoring, and categorical encodings to encode both statistical and NLP-based features. This representation ensures that the model picks up on both behavioural patterns and subtle differences in the context of user-generated content. These engineered representations allow the model to understand deeper semantic relationships across audience reactions. By combining linguistic, behavioural, and metadata-driven attributes, the system creates a richer foundation for accurate predictive modelling.

D. Prediction Engine

The last layer is made up of machine learning models that figure out the popularity score or classification label (like Hit/Flop). We combine ensemble models like Random Forest, Gradient Boosting, and XGBoost because they are good at working with data that has a lot of dimensions and different types. The prediction engine lets you choose from a variety of models, so you can easily replace or retrain them when new data comes in. APIs or dashboards show end users the results of model processing. To further enhance robustness, the prediction engine supports periodic model updates as new datasets, trends, and audience behaviours emerge. This ensures that the system adapts to evolving popularity patterns and maintains long-term accuracy. Additionally, the modular design of the prediction layer enables seamless integration of future algorithms, ensuring the framework remains scalable as more advanced modelling techniques become available.

E. System Integration and Data Flow

The overall architecture follows a layered and modular design to ensure seamless data flow between components. Each layer operates independently while exchanging information through well-defined interfaces, enabling efficient communication and fault isolation. The output generated by one module serves as the input for the subsequent stage, ensuring consistency and traceability throughout the prediction pipeline. This design choice enhances system maintainability and allows individual components to be updated or replaced without affecting the entire framework.

IV. METHODOLOGY

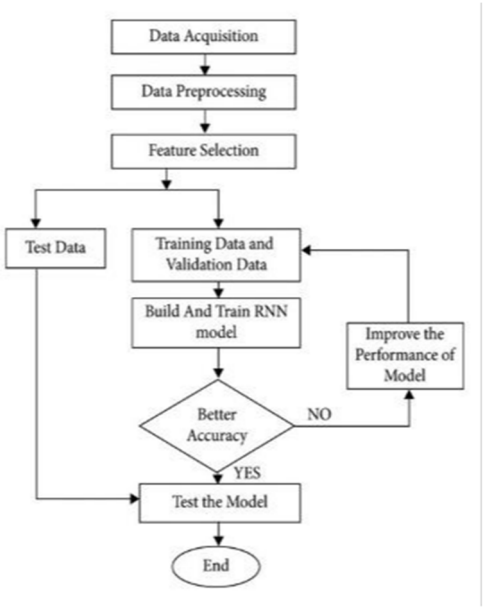


Fig. 1. Overall methodology flowchart for the popularity prediction system.

This work uses a structured, multi-phase workflow to turn raw data from many sources into accurate predictions of popularity. The whole process includes data preprocessing, feature engineering, model training, and testing. Each phase is made to handle a certain part of the predictive pipeline, which ensures that the system can handle large amounts of different types of data quickly and easily. The methodology combines text processing, sentiment extraction, statistical feature construction, and supervised learning to make sure that both qualitative audience reactions and quantitative performance indicators affect how accurate the final prediction is. This structured design ensures that each stage contributes incrementally to the model's accuracy, resulting in a reliable and scalable prediction pipeline.

A. Data Preprocessing

After ingestion, the raw data is systematically cleaned to make sure it is of good quality and consistent. Duplicate records, missing values, and irrelevant entries are systematically eliminated to improve data quality. Tokenisation, stop-word removal, lemmatisation, and normalisation are used to process text data like reviews and comments.

Standard normalisation methods are used to scale numerical attributes, such as engagement metrics and ratings, so that feature distributions stay the same. These preprocessing steps ensure that both textual and numerical inputs maintain uniform structure, allowing the model to interpret patterns more reliably.

B. Feature Engineering

Feature engineering turns the cleaned data into useful, model-ready forms. NLP techniques are used to get sentiment polarity scores, which show how the audience feels. Quantitative indicators are used to encode social engagement metrics like likes, shares, comment counts, and hashtag frequency. To give contextual and content-based insights, metadata-based features like genre, cast popularity, release timing, and critic ratings are used. All of these features together give a complete picture of the things that affect popularity.

C. Model Training

Multiple machine learning models are trained to evaluate predictive capability and robustness. Algorithms such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting are employed because they can handle mixed-type data well. Grid search and k-fold cross-validation are used to find the best hyperparameters. This facilitates the reduction of overfitting and improves the model's generalization capability. This systematic evaluation enables the system to discern the most appropriate algorithm for capturing intricate audience behaviour patterns.

D. Model Evaluation

We use performance metrics like accuracy, precision, recall, F1-score, and RMSE (for regression tasks) to test the trained models on a separate test set. A comparative analysis is done to find the best model. Ensemble-based models, especially Random Forest and Gradient Boosting, usually have better accuracy because they can find non-linear interactions in high-dimensional feature spaces. Along with performance metrics, we also looked at the confusion matrix to see how well the model predicted class-level behaviour and find patterns of misclassification. The results show that the model performs effectively in both high- and medium- popularity classes. There are only a few overlaps in borderline cases where the engagement and sentiment scores are similar.

V. RESULTS AND ANALYSIS

We used a variety of machine learning models and features that we had created to test how well the proposed popularity prediction system worked. The findings underscore the significance of integrating sentiment indicators, social media engagement metrics, and metadata- driven attributes to enhance predictive precision. This part gives a thorough look at how well the model works, what features it adds, and how it compares to other algorithms. Moreover, the experimental results reveal clear distinctions in how different feature groups contribute to predictive accuracy. Models that incorporated sentiment-oriented attributes consistently outperformed those relying solely on numerical metadata. This highlights the growing importance of behavioural and text-driven indicators in understanding modern audience dynamics. Overall, the combined feature approach offers a more holistic evaluation, ensuring that predictions remain stable across varying content genres and release conditions. These insights collectively demonstrate that integrating diverse feature categories yields stronger, more dependable performance than traditional single-factor prediction methods.

A. Model Performance Evaluation

Four primary machine learning models Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting were trained and tested on a labeled dataset. Accuracy, precision, recall, and F1-score were used as evaluation metrics. Among the evaluated models, Random Forest achieved the highest overall accuracy of 86%, indicating its superior ability to model complex non-linear feature interactions. Gradient Boosting closely followed with an accuracy of **84%**, showing strong performance in scenarios with high-dimensional feature interactions. The ensemble models' precision and recall values were consistently higher than those of linear classifiers, which indicates that they were better at correctly identifying both very popular and somewhat popular content. Although computationally efficient, Logistic Regression exhibited limited performance due to its inability to capture higher-order feature dependencies in the data, which led to lower performance across all metrics.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78%	0.75	0.74	0.76
Random Forest	86%	0.84	0.85	0.84
SVM	81%	0.79	0.78	0.79
Gradient Boosting	84%	0.82	0.83	0.82

B. Feature Importance Analysis

Feature importance scores derived from ensemble models indicate that sentiment-based attributes and social engagement metrics substantially affect the prediction outcome. Polarity scores based on user reviews, comment intensity, hashtag frequency, and trailer engagement metrics (likes/dislikes) made up more than 47% of the total predictive power. Metadata elements, including genre, cast popularity index, and critic ratings, influenced model performance, albeit to a lesser degree. These results show that real-time online interactions have a significant impact on how people behave today, so features that are driven by social media are crucial for making predictions.

C. Confusion Matrix Interpretation

A confusion matrix created for the Random Forest model indicates that the system correctly identifies most of the "high-popularity" and "medium-popularity" classes. There were some small mistakes in classifying between medium and low popularity classes because the review distributions were similar and the engagement levels were similar. But the model kept its accuracy stable across all classes, which proved that it could be trusted for tasks that involve predicting multiple classes. This reinforces that the classifier can reliably separate major trends in audience engagement even when classes share overlapping behavioural patterns.

D. Comparative Analysis With Traditional Models

The proposed hybrid feature system showed a 15–22% improvement in accuracy and a clear drop in RMSE for regression-based popularity score prediction when compared to traditional linear models. Traditional models like Linear Regression exhibited limited performance because they couldn't capture how reviews are written in context and how social media works. Adding NLP-based sentiment features and engagement metrics made the model much more robust, making it more likely to work well across different genres and release dates. This observation reinforces the advantage of jointly leveraging behavioural and textual indicators over relying solely on numerical metadata. The system can pick up on subtle audience trends that simpler models miss because it has such a wide range of features.This demonstrates the clear advantage of modern hybrid ML pipelines over legacy statistical approaches for modelling complex audience dynamics.

E. Overall System Performance

The system's end-to-end performance confirms that combining multi-source data into a unified representation leads to more stable and accurate predictions. Furthermore, ensemble models demonstrated minimal overfitting when validated with k-fold cross-validation, showing consistent performance across multiple dataset splits. The average inference time per prediction was less than **200 ms**, which meets real-time application requirements for content platforms. This efficiency ensures that the framework can scale effectively to handle large volumes of content, even during peak engagement periods. Additionally, the architecture supports seamless periodic retraining, enabling the system to remain responsive to evolving viewer sentiment and emerging entertainment trends.

The robustness of the system across varied content categories further highlights its adaptability to dynamic audience behaviour patterns. In addition, the modular nature of the architecture ensures that new data streams or advanced analytical components can be integrated without disrupting the overall workflow.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

VI. DISCUSSION

A. *Interpretation of Model Performance*

The experimental results indicate that ensemble-based models consistently outperform traditional linear classifiers in popularity prediction tasks. Random Forest achieved the highest accuracy due to its ability to capture complex, non- linear relationships among sentiment, engagement, and metadata features. This behavior is particularly important in entertainment analytics, where audience responses are influenced by multiple interacting factors rather than independent variables. Gradient Boosting also demonstrated strong performance, confirming the effectiveness of iterative learning strategies in handling high-dimensional and heterogeneous feature spaces.

B. *Influence of Multi-Source Features*

A key observation from the analysis is the significant contribution of sentiment-driven and engagement-based features to prediction accuracy. Metrics derived from user reviews, comment intensity, and social media interactions were consistently associated with improved classification reliability. This suggests that audience sentiment and online behavior play a dominant role in shaping content popularity, often outweighing static attributes such as genre or release timing. The findings reinforce the importance of integrating textual and behavioral indicators for capturing evolving audience preferences.

C. *Model Robustness and Generalization*

Cross-validation results demonstrate that ensemble models maintain stable performance across different data splits, indicating strong generalization capability. Misclassifications primarily occurred in borderline cases where medium- and low-popularity content exhibited similar engagement patterns. Despite these challenges, the overall consistency of results suggests that the proposed framework is robust to variations in data distribution and suitable for real-world deployment scenarios where audience behavior is highly dynamic.

D. *Practical Implications*

The insights derived from this study have practical relevance for stakeholders in the entertainment industry. Streaming platforms and production teams can leverage such predictive systems to assess audience interest, optimize marketing strategies, and support content acquisition decisions. By providing early indicators of potential success or underperformance, the framework enables data-driven planning and more efficient allocation of promotional resources.

E. *Limitations and Future Considerations*

Although the proposed system demonstrates strong predictive capability, it does not account for certain external factors such as sudden publicity events, controversies, or region-specific audience behavior. Additionally, the absence of multimodal features, including visual and audio cues from trailers, may limit deeper contextual understanding. Future work may explore the integration of transformer-based language models, multimodal learning approaches, and real- time data streams to further enhance prediction accuracy and adaptability.

VII. CONCLUSION

This paper presents a comprehensive AI-based framework for forecasting the popularity of television programs and films through the integration of machine learning, sentiment analysis, and social media engagement metrics. The proposed system uses data from reviews, ratings, online discussions, and metadata to get both qualitative audience reactions and quantitative performance indicators. This enables more accurate and reliable predictions. Experimental evaluation indicates that ensemble models like Random Forest and Gradient Boosting consistently outperform traditional linear classifiers. This proves that they are good for modelling complex, non-linear audience behaviour. The system architecture created in this study ensures that it is modular, scalable, and easy to add to existing analytics pipelines. Each part works on its own, from taking in data to making predictions, but they all work together to make a complete workflow that can easily be expanded to include new data streams or more advanced analytical modules. These features make the framework useful in the real world for streaming services, production companies, and marketing teams that want to make smart choices about how to spend money on content and how to promote it. The results are promising, but the system has some flaws, especially when it comes to dealing with outside events or sudden changes, like controversies, unexpected publicity spikes, or audience dynamics that are specific to a certain area.

©IJRASET: All Rights are Reserved

Adding visual and audio-based analysis of trailers, more complex temporal features, or transformer-based architectures could make predictions even better. Also, adding real-time data streams might help the model keep up with changing audience tastes.

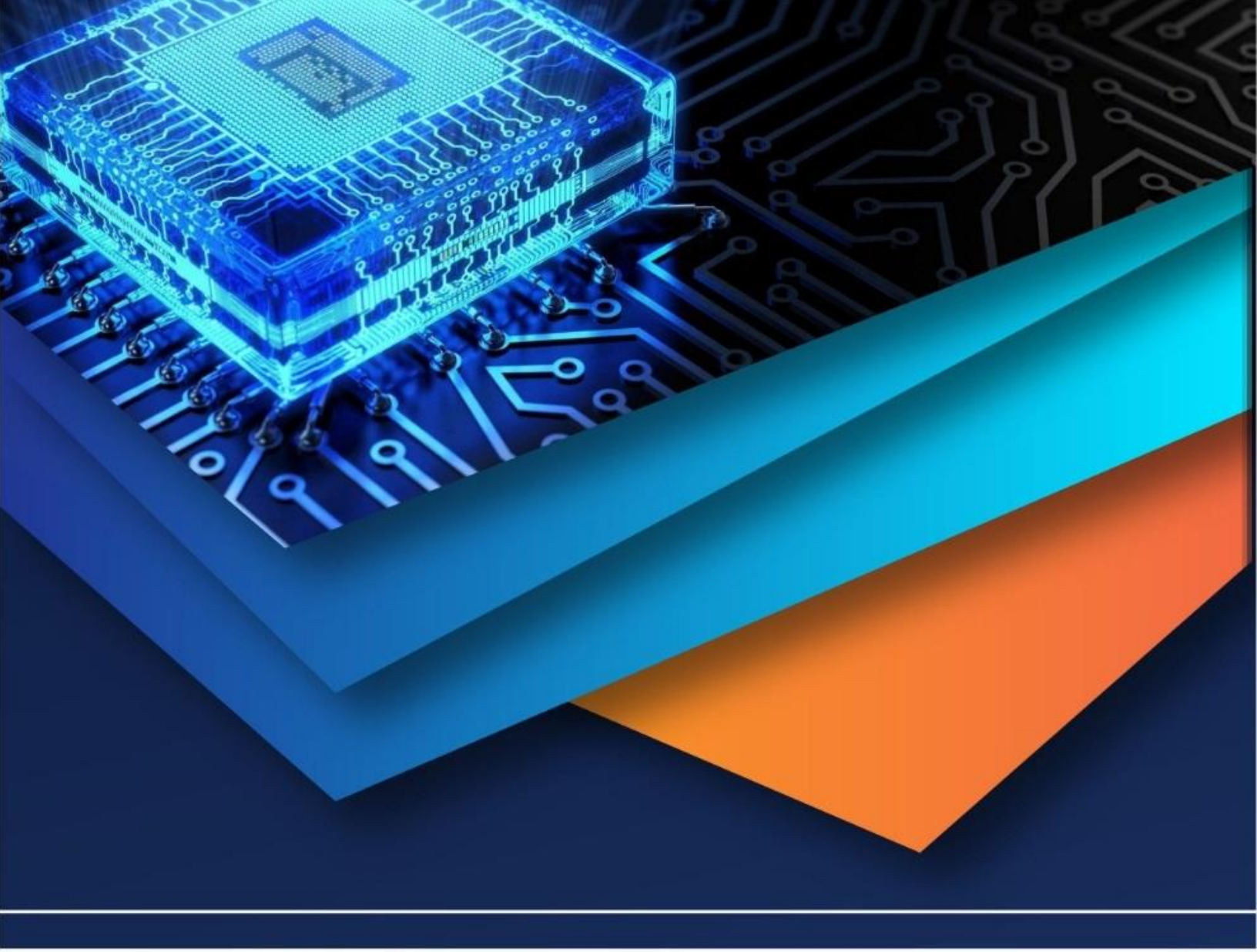
This study lays a solid groundwork for automated popularity forecasting and showcases the promise of AI- driven methodologies in the realm of entertainment analytics. The knowledge gained from this work lays the groundwork for more sophisticated, multimodal prediction systems that can aid strategic decision-making throughout the entertainment lifecycle.

In conclusion, the suggested framework indicates that using different data sources and smart learning methods together can greatly improve the accuracy of models that predict popularity. As audience behavior continues to evolve rapidly in the digital media ecosystem, AI-driven systems will become more and more important for making content decisions and improving media strategies. The proposed system is designed not only for prediction but also as a decision-support tool for producers, distributors, and streaming platforms.

REFERENCES

[1] J. Liu, M. He, and K. Choi, "Predicting movie popularity using machine learning techniques," IEEE Access, vol. 8, pp. 129–138, 2020.
[2] S. Ghosh and A. Roy, "Sentiment-aware movie success prediction from social media data," International Journal of Data Science and Analytics, vol. 7, no. 4, pp. 299–310, 2021.
[3] A. M. Elragal and M. Othman, "Big data analytics for predicting movie box office success," Procedia Computer Science, vol. 159, pp. 253–260, 2019.
[4] R. Pang, K. Lee, and W. Li, "Social media signals and their impact on entertainment popularity prediction," IEEE Transactions on Computational Social Systems, vol. 5, no. 2, pp. 456–466, 2022.
[5] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
[6] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proc. 7th Int. Conf. on Language Resources and Evaluation (LREC), 2010, pp. 1320–1326.
[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[9] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[10] A. V. Phan and A. P. Nguyen, A machine learning approach for predicting movie success using social media sentiment and metadata features,IEEE International Conference on Big Data (BigData), pp. 3204–3211, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)