Movie Schedule Optimization using Sentiment Analysis of Movie Reviews

Dr. S P Siddique Ibrahim
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
siddique.ibrahim@vitap.ac.in

A. Mohan Krishna
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
mohan.21mis7128@vitapstudent.ac.in

P. Saiva Sri
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
saiva.21mis7163@vitapstudent.ac.in

D. Ragha Sri Pranati
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
pranati.21mis7161@vitapstudent.ac.in

P. Sandeep Kumar
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
sandeep.21mis7158@vitapstudent.ac.in

R. Sai Vighnesh
School of Computer Science and Engineering
VIT-AP University
Amaravathi, India
vighnesh.21mis7157@vitapstudent.ac.in

ABSTRACT

Within the advanced film industry, distributors frequently facing challenges in ideally assigning theaters and showtimes for recently released movies. Conventional assignment strategies can result in effective movies getting less theaters than required, whereas normal movies possess more space than justified, driving to disappointment among movie lovers and budgetary misfortunes for distributors. To solve this, we propose a Movie Schedule Optimization using Sentiment Analysis. It is by analyzing first-day movie reviews by utilizing algorithms such as Logistic Regression, SVM, K-Nearest Neighbors(KNN), and Naïve Bayes, we are able predict the sentiment of a film. Based on this examination, we prescribe alterations within the number of theaters for each film. The Multinomial Naive Bayes show, appearing the most elevated precision, is chosen as the ultimate demonstrate for this errand. Our approach points to progress group of onlookers fulfillment and minimize distributor misfortunes by powerfully altering movie plans based on real-time estimation examination.

KEYWORDS

Polarity, Subjectivity, Reviews, KNN, Multinomial Naive Bayes

1. INTRODUCTION

The triumph of a movie's theatrical run is greatly depends on the selection of theaters and showtimes. But current methods frequently overlook audience reactions in real time, which results in an ineffective use of resources. Many popular movies have trouble finding theaters to screen them, and too many movies with bad reviews end up in theaters, frustrating moviegoers and hurting distributors' bottom lines.

Our study uses sentiment analysis on first-day movie reviews to examine a fresh solution to this problem. By compiling and analyzing these reviews, we can predict how well or poorly a film will do at theaters. This data-driven approach may allow us to recommend adjustments to the number of theaters and showtimes for each movie.In order to categorize the tone of evaluations, we utilize several machine learningTo classify the sentiment of reviews, we employ a number of machine learning methods, including SVM, Logistic Regression, Naïve Bayes and KNN.

The model Multinomial Naive Bayes is used for final recommendations because of its greater accuracy. Our goal in incorporating sentiment analysis into the decision-making process is to make movie scheduling more effective. This would not only help distributors cut losses but also improve audience demand. This concept's dynamic and adaptable venue allocation mechanism benefits both moviegoers and the film distribution business.

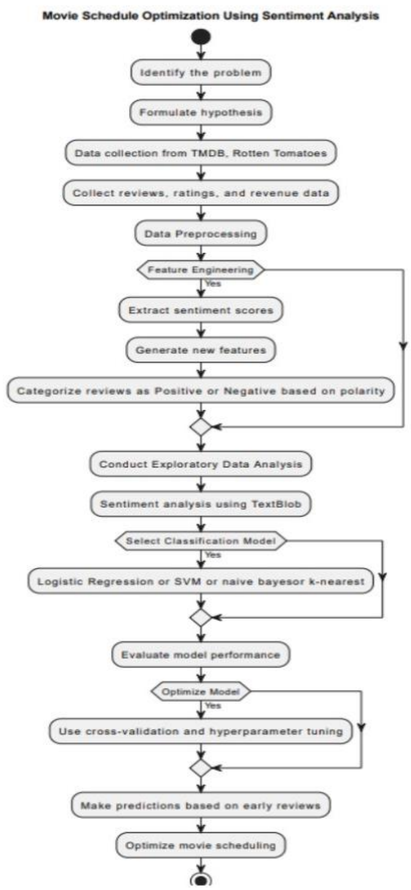

Fig - 1

2. LITERATURE REVIEW

Sentiment analysis, also called as review mining, is widely used to determine the sentiment or emotional tone of text data. Generally movie reviews are positive or negative which can be classified by Sentiment analysis, which is useful for understanding audience feedback. The papers explore different supervised machine learning methods, such as Random Forest, SVM, Naïve Bayes and KNN in the context of sentiment analysis.

Naive Bayes is a probabilistic model that is popular for text classification due to its simplicity and efficiency. Paper 3 used Multinomial Naive Bayes (MNNB) in combination with Gini Index for feature selection, showing a slight improvement in accuracy (from 56% to 59.54%) when using feature selection [1].

k-Nearest Neighbors (KNN) follows finding the Nearest data points in the feature space and is effective for classification tasks. Papers 1 and 2 compared k-NN with other models and found that while it can perform well, it often lags behind other more sophisticated algorithms in accuracy [3][2].

Support Vector Machine (SVM) is known for its classification tasks. Paper 2 found that SVM outperformed KNN and Naïve Bayes for sentiment classification, achieving accuracies of at least 80%, especially when the training dataset size increased beyond 150 reviews[2]. This paper investigates a machine learning-based approach to opinion examination by applying support vector machines (SVM) and outfit learning. The creators utilize the chi-square selector for include determination and tune SVM hyperparameters with GridSearch. Furthermore, they upgrade the execution employing a sacking calculation. Key discoveries appear that this combination outflanks conventional estimation examination models in classifying motion picture surveys. [4]

Random Forest, a group learning method, was also used for sentiment analysis. It aggregates the output of several decision trees, improving accuracy and generalizability. In Paper 1, Random Forest was included among the classifiers, showing promising performance [1].

Long Short Term Memory(LSTM) systems, a type of Recurrent Neural Network (RNN), for opinion examination of motion picture audits. The creators contend that conventional models like SVM and direct regression confront confinements when managing with long literary groupings due to issues such as vanishing slopes. LSTMs, with their capacity to preserve long-term conditions, are proposed as a prevalent elective for preparing long surveys. The demonstrate is prepared on the IMDB dataset, and word embeddings like Word2Vec are utilized for highlight representation. The study appears that LSTMs outperform conventional models in capturing sentiment extremity, but the creators too emphasize the significance of hyperparameter tuning (e.g., dropout rates, number of layers) for optimal performance. [6]

Lexicon-based sentiment analysis relies on dictionaries that classify words as positive or negative. By building a lexicon of words and phrases, sentiment is derived from the presence or absence of positive or negative terms. This method is often paired with techniques like Part of Speech tagging (POS) and Name Entity Recognition (NER) to improve results. Some techniques are Latent Dirichlet Allocation (LDA) and Term Frequency-Inverse Document Frequency(TF-IDF) are applied to enhance feature extraction [5].

3. METHODOLOGY

3.1 K-Nearest Neighbor

k-Nearest Neighbor (KNN) is a supervised, memory-based learning calculation utilized for both classification and relapse assignments. KNN makes forecasts by finding the k preparing illustrations that are closest to a new data point (based on a distance metric like Euclidean remove) and after that making a prediction based on their values. For classification, it allots the lesson that's most common among the k neighbors.

KNN Steps:

i. Calculate the remove between the test occasion and all preparing cases.
ii. Select the k illustrations with the littlest remove.
iii. For classification, take a larger part vote among the k neighbors to choose the course name. For relapse, take the normal of the k neighbors.

3.2 Logistic Regression

Logistic Regression could be a factual show utilized for binary classification issues, which predicts the likelihood of an instance belonging to one of two classes. Rather than fitting a line like in linear regression, logistic regression employments the sigmoid work to deliver a likelihood yield that's between 0 and 1. It finds the best-fitting demonstrate to depict the relationship between a target variable and set of highlights.

$$P(y = 1/x) = \frac{1}{1 + e^{-z}}$$

Logistic regression is widely used for tasks like spam detection, disease prediction, and credit scoring due to its simplicity and interpretability.

3.3 Support Vector Machine

Support Vector Machine (SVM) is a well-defined learning calculation utilized for both classification and regression tasks. SVM points to find the ideal hyperplane that maximally seperates information focuses of distinctive classes. For a parallel classification, this hyperplane is chosen such that the edge between the closest information focuses from both classes is maximized. SVM can utilize bit capacities to convert non-linearly distinct information into a higher-dimensional space, where a hyperplane can effectively classify the information. Common bit capacities incorporate straight, polynomial, and Radial Basis Function (RBF) parts.

3.4 Naïve Bayes

Naive Bayes may be a basic but classification calculation based on Baye's Hypothesis. It is especially well known for content classification issues, such as separation of spams and assumption investigation. The "naive" portion of the title suggest to the suspicion that all highlights are free of each other, which is once in a while genuine in real-world information, but it makes the computation much less difficult.

$$P(c|x) = \frac{P(x|c) \cdot P(c)}{P(x)}$$

Naive Bayes classifies a new instance by calculating the posterior probability for every possible class and then assigns the highest probability label.

1. Data Collection Sources: Rotten Tomatoes and TMDB (The Movie Database) were our two number one resources of statistics. Rotten Tomatoes gave evaluations from specialists and audiences, whereas TMDB supplied data on movie scores, income, genres, and other functions. From 2000 to 2020, a vast spectrum of film genres and viewer comments turned into protected with the aid of the dataset.

Data Size: More than 59,000 rows containing a lot of variables, which include evaluation feedback, earnings, genre, and film rankings, have been processed.

2. Data Preprocessing: In order to make sure the dataset's satisfactory, rows with null or missing values had been removed.

Feature Engineering: Using TextBlob for sentiment analysis, we generated functions. In order to provide sentiment labels, evaluation scores were additionally divided into "Fresh" and "Rotten" classes relying on their ratings.

Data Encoding: Numerical values were assigned to categorical variables, along with review categories and scores. A "Fresh" evaluation, as an instance, was encoded with a fantastic sentiment of 1 and a "Rotten" overview with a terrible sentiment of zero.

3.Exploratory records Analysis (EDA):

We performed exploratory records analysis to look how ratings numerous between assessment ranks. This made it easier to discover tendencies linking favorable opinions to better movie rankings.

Distribution of Polarity: Higher assessment scores (A or B ratings) were more likely to have nice polarity scores, at the same time as decrease scores had been more likely to have terrible polarity ratings.

Link Analysis: Using a linear regression version, we were able to show a high quality hyperlink among the polarity score and the review score. Review rankings were usually more for films that had greater desirable evaluations, or better polarity.

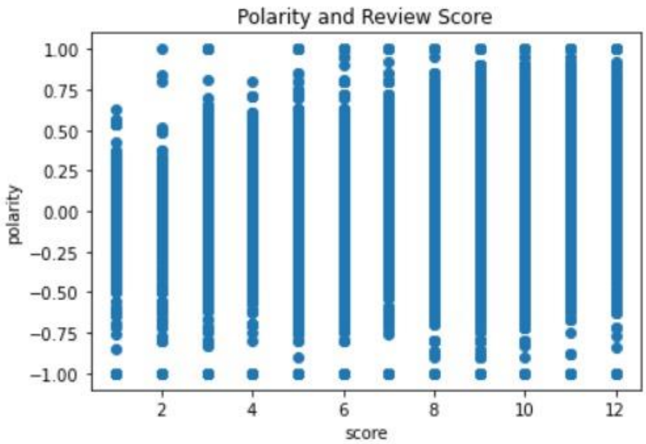

Fig - 2

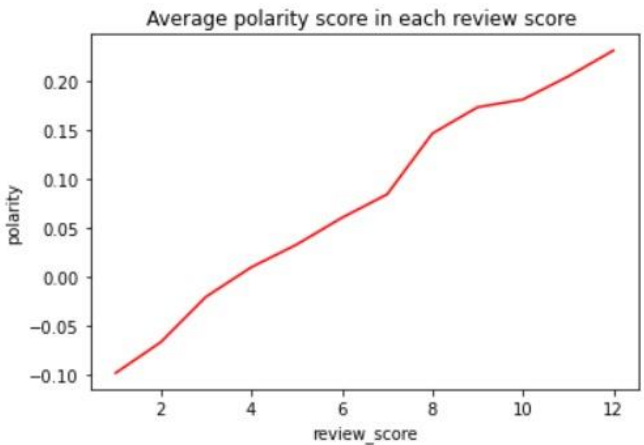

Fig - 3

4. Sentimental Analysis:

Sentiment Extraction: The TextBlob software changed into utilised to extract sentiment from textual comments. Subjectivity assessed the degree to which a remark changed into goal or subjective, and polarity scores ranged from -1 (poor) to one (fantastic) for each announcement.

Relationship between Polarity and Subjectivity: We discovered that extra excessive remarks had been additionally greater subjective, assisting our speculation that extraordinarily subjective emotional critiques have been.

5.Predictive Modelling:
Setting Up for Models: Two strategies were employed to tokenize the assessment contents and transform them into numerical vectors:
Count Vectorizer: This technique took word frequency inside the opinions under consideration.
TF-IDF Vectorizer: This approach took under consideration both word frequency and the frequency with which a term appeared in various critiques.

6. Model Evaluation and Performance:
Measures of Accuracy: We tested the accuracy of the TF-IDF and depend vectorizer techniques. On the check set, the Multinomial Naive Bayes version with the matter vectorizer yielded the nice accuracy.
Confusion Chart: The Naive Bayes version produced balanced predictions for each tremendous and terrible sentiments, unlike the other fashions, which tended to over-expect superb sentiments.
Creation of Word Clouds: We created phrase clouds to expose the most usually used effective and bad phrases in evaluations, providing insights at the language that affects the overall impression of these comments.

Table – 1

Accuracy	**Count vectorizer for bag of words (BOW)**		**Tfidf Vectorizer**	
	Test accuracy rate	**Training accuracy rate**	**Test accuracy rate**	**Training accuracy rate**
Logistic Regression	0.6479	0.9322	0.6458	0.6576
SVM	0.6458	0.8618	0.6458	0.6577
Naive Baye's (Multinomial NB)	0.6568	0.9338	0.6473	0.9338
KNN	0.6461	0.6577	0.6458	0.6577

7. Scheduling Optimisation for Movies:
Based on the sentiment analysis, we employed this records to beautify movie scheduling. Key findings from the examination made the subsequent possible:
Predicting Box Office Success: Theatres may also predict a film's overall performance on the field office via inspecting early opinions from releases or check screenings.
Customizing Screening Times: The sentiment analysis assisted theatres in selecting how regularly and when to preserve screenings. Films with a greater optimistic tone and more polarity were scheduled at more prominent instances and venues.
Demographic Insights: Cinemas had been able to maximize income by using choosing places with sturdy target audience interest via tailoring their scheduling to certain target audience agencies by using combining sentiment analysis with demographic records.

8. Commercial Utilisation:
This technique provides a device for film theatre owners and traders to make higher judgements. Stakeholders may prevent field workplace screw ups and maximise income for films with promise by way of assessing early attitudes and selecting the pleasant times and locations for film screenings.
By making use of real-time target audience sentiment data, this technique of combining sentiment analysis, machine mastering models, and linear regression efficaciously tackles the trouble of film agenda optimisation.

4.RESULTS AND DISCUSSION

The study included the utilize of numerous machine learning classifiers—KNN, Logistic Regression, SVM, Multinomial Naïve Bayes(MNB) to classify content information based on two distinctive include extraction strategies:
TF-IDF Vectorizer and Bag of Words (BOW). The test and preparing exactness for each combination of demonstrate and include extraction strategy are summarized in Table 1.

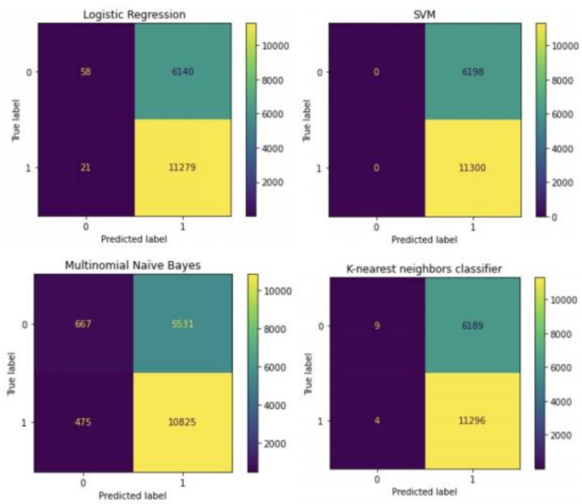

Fig - 4

Confusion Matrix Analysis
The confusion matrices give extra insights into the classification performance of each model. The true label and predicted label are plotted to better get it the distribution of accurately and inaccurately classified occurrences.
Logistic Regression:
It is obvious from the confusion matrix that Logistic Regression had a critical number of untrue positives and untrue negatives, showing that the demonstrate struggled to precisely anticipate minority class labels.
SVM:
This model showed no false positives and was more reliable in anticipating the majority class label, in spite

of the fact that it missed predicting minority class labels inside and out.

Naïve Bayes:

The Multinomial Naïve Bayes classifier had moderately balanced performance, in spite of the fact that it delivered a few false positives and false negatives. The model shows up to have had more trouble recognizing between true positives and false negatives, driving to marginally lower precision.

K-Nearest Neighbors (KNN):

KNN had a little number of untrue positives and a better number of untrue negatives compared to SVM, demonstrating that it was not as viable at accurately foreseeing the minority class.

5. CONCLUSION

An inventive and data-driven strategy for comprehending group of onlookers inclinations and moving forward theater operations is given by the utilize of opinion investigation into motion picture plan enhancement. Ordinary motion picture planning strategies habitually put a extraordinary bargain of weight on box office comes about, past execution designs, or audits from critics—all of which may not precisely reflect how an gathering of people feels almost a picture. Cinema administrators can make superior choices by utilizing assumption examination, which is determined from the handling of motion picture surveys and offers a more nuanced viewpoint on how the open sees a film. This strategy ensures that movies that positively interface with watchers are given priority in prime openings by enhancing the plan based on real-time estimation information, which seem increment income creation and boost client happiness. The recommended approach organizes making strides moviegoing encounters in expansion to benefit boost.

Cinemas may provide a more customized seeing encounter that increments client consistency by adjusting the development picture orchestrate based on the energetic and subjective input of the swarm. Other than, estimation examination gives a flexible approach that lets chairmen change plans in response to startling bunch of on lookers responses or changes in open supposition. Much obliged to this responsiveness, theaters may advantage from unexpected spikes in ask for particular motion pictures while decreasing the dangers related with awful scheduling. Overall, this consider outlines that suspicion examination can change the schedule approaches to movement picture arranging into a appear that's more versatile, flexible, and centered on needs of the customer. It makes it conceivable to change client elation with corporate progression. To ensure careful organizing techniques, assumption examination got to be utilized in concert with other contraptions for decision-making, in reality in appear abhor toward of the truth that it offers sharp data on its claim. This set of contraptions gives a novel way to advance gathering of people seeing encounters and boost box office execution for theaters.

6. FUTURE SCOPE

There are a few ways to carry out more inquire about and move forward assumption analysis's utilize in motion picture planning advancement. Utilizing more complex normal dialect preparing (NLP) models, such transformers and profound learning calculations, is one vital slant. These models have the potential to improve assumption classification exactness by consolidating more complex phonetic highlights, like setting, mockery, or multi-layered enthusiastic responses in audits. Moreover, the integration of multilingual assumption examination can encourage the understanding of gathering of people opinion over dialects and societies by theaters working in numerous locales, driving to more comprehensive and focused on scheduling.For a more careful and all-encompassing approach, more information sources can be consolidated into the planning handle in expansion to estimation investigation of motion picture audits. For case, real-time bits of knowledge into gathering of people inclinations and the buzz encompassing as of late discharged or future movies can be gotten from social media stage information. Moreover, motion picture theater administrators may tailor their plans to target socioeconomics counting age bunches, sexual orientation inclinations, and geographical trends by coordination statistic information. These variables may well be utilized with assumption investigation to form a more precise and redone planning calculation that would boost box office execution and group of onlookers engagement.One more thing to see into is making real-time, versatile motion picture planning frameworks. The screening plan can be powerfully modified by these calculations by routinely dissecting new gathering of people reaction, estimation shifts, and box office information. For occasion, a film can be changed to a prime-time space inside days in the event that it out of the blue gets awesome surveys or gets to be viral on social media, guaranteeing that theaters respond quickly to client request. This sentiment-based planning methodology seem too be utilized in other excitement businesses, such sports, live occasions, and gushing administrations, having a greater impact in areas where group of onlookers endorsement and happiness are critical victory components.

7. REFERENCES

[1] Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. International Journal of Computer Applications, 179(7). https://doi.org/10.5120/ijca2017916005

[2] Kalaivani, P., & Shunmuganathan, K. L. (2013). Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches. Indian Journal of Computer Science and Engineering, 4(4), 285-290.

[3] Purnomoputra, R. B., Adiwijaya, & Wisesty, U. N. (2019). Sentiment Analysis of Movie Reviews using Naïve Bayes Method with Gini Index Feature Selection. Journal of Data Science and Its Applications, 2(2), 85-94. https://doi.org/10.34818/JDSA.2019.2.36

[4] Sulthana, R. A., Jaithunbi, A. K., Harikrishnan, H., & Varadarajan, V. (2022). Sentiment analysis on movie reviews dataset using support vector machines and ensemble learning. International Journal of Information Technology and Web Engineering, 17(1). https://doi.org/10.4018/IJITWE.311428.

[5] Kulkarni, E. G., & Pujar, A. M. (2021). Intelligent framework for sentiment analysis of movie reviews - A comprehensive survey. Journal of Emerging Technologies and Innovative Research, 8(6), 738-742. http://www.jetir.org/papers/JETIR2106802.pdf

[6] Bodapati, J. D., Veeranjaneyulu, N., & Shaik, S. (2019). Sentiment analysis from movie reviews using LSTMs. Ingenierie des Systemes d'Information, 24(1), 125-129. https://doi.org/10.18280/isi.240119

[7] Lee, & Pang, B. (2008). sentiment analysis as well as opinion mining. Principles and Developments in Information Retrieval, 2(1-2), 1–135.
(doi.org/10.1561/1500000011)

[8] Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. https://doi.org/10.1002/widm.1253

[9] Asur, S. and Huberman, B. A. (2010). using social media to make future predictions. The 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Proceedings, 492-499.
doi.org/10.1109/WI-IAT.2010.63

[10] In 2003, Hatzivassiloglou, V., and Yu, H. Identifying the polarity of opinion sentences and separating opinions from facts are two steps towards providing an opinion-based answer. Empirical Methods in Natural Language Processing (EMNLP) Conference Proceedings, 2003, 129–136. [W03-1017] available at https://www.aclweb.org/anthology

[11] In 2010, Tumasjan, A., Welpe, I. M., Sandner, P. G., and Sprenger, T. O. Predicting elections with Twitter: What 140 characters convey about political mood. The Fourth International AAAI Conference on Social Media and Weblogs, Proceedings, 178–185.
This article can be viewed at https://ojs.aaai.org/index.php/ICWSM/11001