Prediction of Critical Review Scores Using NLP Sentiment Analysis and Film Metadata

Cam Hayes
University of North Florida
n00662220@unf.edu

Abstract

Online film criticism has grown exponentially in the last decade with the growth in popularity of platforms such as Rotten Tomatoes, X, Reddit, and other social media. There are many ways for users of these platforms to express their thoughts and feelings about films, and the effect of that expression can greatly impact box-office performance. Brief, summarized text reviews dominate this space and provide unique challenges and opportunities for analyzing critical opinion at scale. Extracting information from these reviews through techniques such as sentiment analysis provides valuable insight into the critical performance of films through online trends. Sentiment analysis is a language processing technique that attributes emotional weight to a text through varying categories of sentiment. These categories often describe emotions as "positive" or "negative", but advances have been made in models to predict more nuanced emotions such as "joy", and "surprise". However, prediction using sentiment remains a challenge for models that struggle to derive value from texts that are unstructured, indirect, and contradictory, such as those found in online film reviews. This paper proposes a multi-task technique for improving the prediction of critical reviews through a combination of sentiment analysis modeling and film metadata.

1. Introduction

The expression of opinion has exploded alongside the rise of social media, which has become a dominant outlet for sharing our thoughts, especially about products and entertainment. Platforms like Rotten Tomatoes, Reddit, and X have lowered the barriers to entry for film criticism, decentralizing it from traditional critics. As a result, modern online reviews tend to be brief, sarcastic, and tailored to platform-specific audiences. Algorithmic content curation further reinforces selective expression, encouraging focused commentary over holistic film analysis.

The value in sentiment analysis on social media is increasingly significant in industry. Sentiment analysis is a natural language processing technique for extracting emotional content within text, often labeling expressions as "positive," "negative," or more granular emotions like "joy" or "anger." In industries such as entertainment and marketing, sentiment analysis helps extract consumer insights from large volumes of user feedback. In the motion picture industry, pre-release sentiment is sometimes inversely related to box office success — a phenomenon known as the harbinger of failure (Wiles et al. 2023). So, if a production company were able to accurately predict sentiment based on shared criticism, it would be advantageous to respond to changes in marketing and investment strategies. However, sentiment models often struggle (IMB et al. n.d.) to handle sarcasm, context-specific language, and indirect expression, limiting their ability to interpret nuanced opinions in unstructured, user-generated reviews. Because of this, individual sentiment may not always be an accurate indicator of the intrinsic value of a product.

Studying film review scores against sentiment provides a better understanding of how to predict complex opinions over short, targeted, and stylistic texts. This study explores whether combining structured metadata (e.g., genre, director) with unstructured sentiment data can improve review score prediction. In the context of film reviews, sentiment alone can mislead when detached from genre or stylistic intent — for instance, "disturbing" or "grotesque" may imply high praise for a horror film. By incorporating metadata, we give models the context they need to disambiguate tone and intent.

To explore this, multiple sentiment models were evaluated. Additionally, a custom regression model for sentiment analysis was trained based on critical text, and several predictive models were tested using both sentiment-only and sentiment+metadata feature sets. This multi-stage experiment compares the strengths and weaknesses of different modeling strategies in a stylized, online review environment.

2. Related work

2.1. The Effect of Sarcasm on Natural Language Processing

A persistent challenge in sentiment analysis is the accurate interpretation of sarcasm and double meaning, which can result in significant misclassification. Prior research highlights the limitations of traditional models in handling such edge cases (Tayyar et al. 2023). Recent work has shown that context-aware deep learning architectures—particularly transformer-based models—are more effective at detecting sarcastic or ironic tone, leading to improved sentiment scoring in informal or stylized text.

2.2. Emotion Classification for Models

Efforts like Google's GoEmotions dataset (Demszky et al. 2021) have advanced sentiment analysis by moving beyond binary classification and introducing fine-grained emotion labels (e.g., "joy," "disappointment," "confusion"). These annotated datasets enable models to encode emotion in a multi-dimensional vector space, supporting more nuanced and interpretable sentiment extraction. Such models offer a promising foundation for domain-specific tasks, including predictive modeling of review scores based on emotional tone.

3. Methodology

This study investigates the relationship between text sentiment and critical review scores using a multi-stage modeling approach. The source data consists of approximately 1 million film reviews and metadata entries derived from Rotten Tomatoes critic reviews and corresponding film information (Leone et al. 2022). These two datasets were merged based on title to associate each review with relevant metadata, such as genre and cast.

Each review was preprocessed and tokenized, preparing the text for sentiment scoring using four sentiment analysis models. The primary modeling tasks in this study are:

1. Sentiment scoring: assigning quantitative emotional values to review text using pre-trained and custom-trained models.
2. Review score prediction: using sentiment features—with and without film metadata—to predict a normalized review score.

A multi-task learning framework is appropriate for this study because sentiment and review score prediction are related tasks, both relying on understanding emotional tone, but only one being supervised directly. By exploring how sentiment-derived features, contextualized with metadata, contribute to score prediction, this study implicitly treats sentiment extraction as a feature-building task in service of a more complex regression model.

3.2 Sentiment Model Evaluation

Four sentiment models were used to extract emotional tone from critic reviews:

- VADER: A lexicon-based model designed for social media sentiment, operating on rules and word polarity scores.
- Distilled RoBERTa: A lightweight transformer-based model derived from RoBERTa and fine-tuned for general sentiment classification.
- Twitter RoBERTa: A variant of RoBERTa trained on sentiment datasets from Twitter posts, more suitable for detecting irony and informal tone.
- FilmBERT: A custom sentiment model developed for this study by fine-tuning the RoBERTa base architecture on Rotten Tomatoes reviews, with review scores as supervision targets.

These models were used to produce numerical sentiment vectors from each review, serving as primary features for score prediction.

3.3 Feature Engineering and Metadata

To improve predictive accuracy and determine the role of context, structured metadata was integrated with unstructured sentiment data. Each review was paired with the following categorical fields:

- Film genre(s)
- Film author(s)
- Film director(s)
- Top-billed actors

Categorical fields were transformed using one-hot encoding. The encoding was limited to the *n*-most frequent values: the top 200 authors, top 200 directors, and top 500 actors across the dataset.

Review scores themselves required standardization. Because critic scores on Rotten Tomatoes appear in multiple formats such as scalar (e.g., "4/5"), letter grades (e.g., "A-"), and ambiguous terms (e.g., "Good", "Worth seeing"), a multi-step normalization process was applied:

- Scalar and letter grades were mapped to a normalized 0–10 scale.
- Ambiguous or indecipherable scores were excluded from the dataset.
- Only reviews with accompanying written text were retained.

This preprocessing produced a dataset with uniform score targets and paired text + metadata features for each review.

3.4 Predictive Modeling Setup

Using the processed dataset, multiple regression models were trained to predict normalized review scores. Each model was evaluated under two experimental conditions:

- Sentiment-only input: models trained exclusively on sentiment vectors derived from the review text.
- Sentiment + metadata input: models trained on sentiment vectors alongside one-hot encoded metadata features.

The models used include:

- Linear Regression
- Ridge Regression
- Random Forest Regressor
- LightGBM
- XGBoost

For each model, hyperparameters were optimized manually to the point at which additional tuning yielded no statistically meaningful improvements in predictive accuracy. This conservative tuning strategy was adopted to reduce the risk of overfitting on a high-dimensional but sparsely populated feature set.

Performance was evaluated using two primary metrics:

- Match Rate (MR): the mean absolute error between predicted and true scores, used to quantify approximate accuracy.
- Coefficient of Determination (R^2): used to assess model ability to explain variance in review scores.

These metrics are presented in Tables 1 and 2, comparing the impact of different sentiment models and the contribution of metadata to prediction accuracy.

4. Experiment

4.1 Predictive Power of Metadata Alone

Early results suggested that metadata features alone could predict general trends in cumulative review scores, albeit with limited precision. Model accuracy was hindered by high variance and noise in the dataset, leading to wide prediction errors despite the presence of an overall linear trend. Nonetheless, the predictive value of metadata was sufficient to justify its inclusion in subsequent experiments alongside sentiment features.

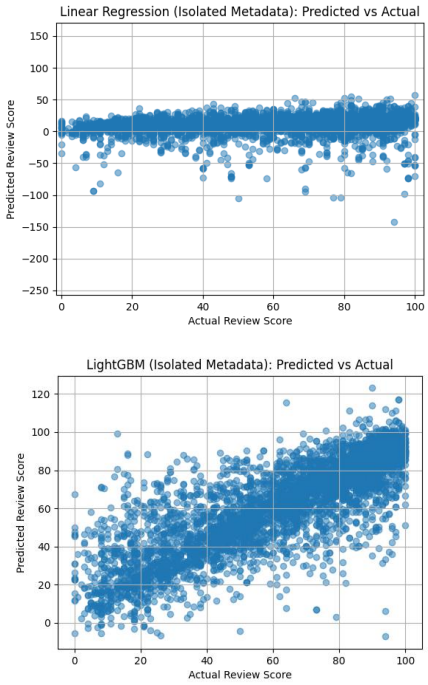

Fig A. Score prediction from encoded metadata, showing non-linear and linear trendlines across noisy data using a linear regression model, and a LightGBM model.

4.2 Custom Sentiment Model: FilmBERT

To test the value of domain-specific sentiment modeling, a custom transformer-based model was trained using the RoBERTa base architecture. This model, named FilmBERT, was fine-tuned to regress review scores directly from review content.

- Architecture: Transformer encoder using RoBERTa-base pretrained weights (Facebook et al. n.d.)
- Optimization: Adam optimizer (Kingma et al. 2015) with learning rate ε=1e-5, α=4, β_1=0.9, β_2=0.98, weight decay = 0.1
- Training Strategy: Epoch evaluation with early stopping based on validation loss

4.3 Predicting Review Scores from Sentiment Features

The predictive regression models were trained on sentiment-derived features to evaluate their accuracy and ability to reason outcomes. Performance was assessed using match rate (mean absolute error between predicted and true review scores) and coefficient of determination R^2, which reflects the proportion of variance explained by the model (Scitkit et al. n.d.).

The results presented below reflect each model's performance across varying sentiment inputs, with and without accompanying metadata.

See Table Legend[TR] in references.

4.3.1 Results: Sentiment Only

	LR	LGBM	RF	RR	XGB
V	**7.8%**	**8.2%**	**8.3%**	**7.8%**	**8.1%**
DR	**8.2%**	**10.4%**	**10.9%**	**8.2%**	**11.6%**
TR	**13.2%**	**9.9%**	**11.5%**	**13.2%**	**11.4%**
FB	**15.2%**	**16.2%**	**16.2%**	**16.2%**	**16.2%**

Table 1.a.: Match rate comparison between regression models on data containing only sentiment features. Match rates indicate a level of precision within 0.25.

	LR	LGBM	RF	RR	XGB
V	**0.05**	**0.04**	**-0.05**	**0.05**	**0.05**
DR	**0.11**	**-0.08**	**0.13**	**0.11**	**0.001**
TR	**0.27**	**-0.04**	**0.17**	**0.27**	**-0.17**
FR	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**

Table 1.b.: Determination between regression models on data containing only sentiment features. Match rates indicate a level of precision within 0.25.

Key Observations:

- FilmBERT outperforms all other sentiment models across all regressors, with a consistent match rate of 16.2% and R^2 of 0.50.
- Twitter RoBERTa outperforms Distilled RoBERTa, while VADER performs worst overall.
- These results demonstrate that context-aware, transformer-based sentiment models provide the most informative signals for predicting critic scores.

4.3.2 Results: Sentiment + Metadata

	LR	LGBM	RF	RR	XGB
V	**12.2%**	**10.8%**	**12%**	**12%**	**12.2%**
DR	**12.6%**	**13.3%**	**14.4%**	**14.4%**	**13%**
TR	**13.5%**	**12.4%**	**14.1%**	**14.1%**	**14.3%**
FB	**15.4%**	**13%**	**15.8%**	**15.8%**	**16.1%**

Table 2.a.: Match rate comparison between regression models on data containing metadata and sentiment features. Match rates indicate a level of precision within 0.25.

	LR	LGBM	RF	RR	XGB
V	**0.36**	**0.09**	**0.34**	**0.34**	**0.23**
DR	**0.38**	**0.25**	**0.12**	**0.12**	**-0.62**
TR	**0.46**	**0.12**	**0.44**	**0.44**	**0.39**
FB	**0.56**	**0.23**	**0.47**	**0.47**	**0.28**

Table 2.b.: Determination between regression models on data containing metadata and sentiment features. Match rates indicate a level of precision within 0.25.

Key Observations:

- Models that underperformed in the sentiment-only setup (e.g., VADER, Distilled RoBERTa) showed significant gains when metadata was included.
- In contrast, FilmBERT and Twitter RoBERTa saw limited or slightly negative gains, suggesting that metadata may contribute to overfitting in high performing models.
- The best overall performance (R^2 = 0.56) came from FilmBERT + Linear Regression, indicating a synergistic effect between contextual sentiment and metadata when paired with a simple regressor.

5. Conclusions

This study set out to evaluate the predictive performance of sentiment analysis models on film critic reviews, both in isolation and in combination with structured metadata. By comparing a range of sentiment models—from rule-based (VADER) to deep contextual models (RoBERTa variants and FilmBERT)—across multiple regression frameworks, the experiments demonstrated that sentiment is a non-trivial prediction metric for review and opinion scoring under appropriately standardized conditions. It was also revealed that context-aware sentiment modeling significantly improves predictive accuracy. Notably, the domain-trained FilmBERT model outperformed all other sentiment models, achieving the highest match rates and R^2 scores across all tested regressors.

Metadata features such as genre, director, and cast further enhanced performance in cases where base models underfit

the data. However, the contribution of those features was generally marginal, or even detrimental, when paired with already high-performing sentiment models. This suggests that the benefit of metadata is contingent on the base strength of the sentiment attribution and may introduce noise or redundancy when context is already captured effectively.

These findings highlight the importance of context in sentiment modeling and the potential of task-specific fine-tuning for predictive applications. In domains like film criticism, where tone, irony, and genre-specific language heavily influence text sentiment, models must be trained not just to detect emotional weight but to understand *how* that weight functions within a stylistic framework.

Future work may involve incorporating fine-grained emotion labels, exploring critic sentiment trends, or testing the model's performance across other entertainment domains such as television or video games. Additionally, the utility of prediction may be assessed further by modifying the acceptable domain of accuracy beyond the parameters defined by this study, and by leveraging standardized scoring systems. Ultimately, this research supports the growing view that hybrid approaches—blending structured metadata with deep, domain-aware language modeling—can yield more accurate and interpretable predictive systems.

6. References

[TR]Legend: V: Vader; DR: Distilled RoBERTa; TR: Twitter RoBERTa; FB; FilmBERT; LR: Linear Regression, LGBM: LightGBM; RF: Random Forest; RR: Ridge Regression; XBG: xGBoost

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Wiles, N. (2023). Negative reviews, positive impact: How early criticism affects box office success. Journal of Marketing Letters. https://link.springer.com/article/10.1007/s11002-023-09665-8

IBM. (n.d.). What is sentiment analysis? Retrieved from https://www.ibm.com/think/topics/sentiment-analysis

Tayyar Madabushi, H., & Lee, M. (2023). Sentiment Analysis and Sarcasm Detection Using Deep Multi-Task Learning. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC9985100/

Demszky, D., Movshovitz-Attias, D., Ko, J., et al. (2021). GoEmotions: A Dataset for Fine-Grained Emotion Classification. Google AI Blog. https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/

Facebook AI. (n.d.). roberta-base. Hugging Face. https://huggingface.co/FacebookAI/roberta-base

Kingma, D., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

Leone, S. (2022). Rotten Tomatoes Movies and Critic Reviews Dataset. Kaggle. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Scikit-learn Developers. (n.d.). sklearn.metrics.r2_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

OpenAI. ChatGPT (April 2025). Used for research assistance, language refinement, and formatting guidance. https://chat.openai.com