Jakub Czabok

Università di Bologna

15.02.2025

Sentiment Analysis and Predictive Modeling of Movie Reviews: Exploring Word Trends and Ratings

INTRODUCTION

In today's world, movie reviews written by critics play an important role in shaping how people see films. These reviews are often published on websites like Rotten Tomatoes, Metacritic, and IMDb, and they can influence whether people decide to watch a movie or not. By analyzing the language and emotions in these reviews, we can better understand what makes a movie successful or unsuccessful. This study focuses on using sentiment analysis and text mining to explore the connection between the words used by critics and the average ratings of movies.

The main goal of this research is to analyze the sentiment and the most common words in movie reviews written by critics. I will focus on specific genres and directors to see if certain words or phrases are linked to higher or lower ratings. Additionally, I will build a predictive model to estimate a movie's average rating based on the review text and other factors like genre and director. This research can help filmmakers, critics, and movie platforms better understand what audiences like and how reviews impact a movie's success.

Movie reviews are not just opinions—they can have a big impact on a movie's popularity. For filmmakers, understanding what critics say can help them improve their work. For movie platforms, analyzing reviews can help create better recommendation systems. This study is also important because it shows how text analysis and machine learning can be used to predict movie ratings, which can be useful for marketing and decision-making.

In the literature, two studies stand out for their contributions to the application of sentiment analysis in the movie domain. Rai and Mewada (2017) demonstrated the effectiveness of machine learning approaches in classifying movie reviews into positive and negative categories, thereby highlighting the potential of textual analysis to extract meaningful sentiment information from reviews. Complementing this work, Zhang, Skiena, and Letchford (2019) investigated how sentiment derived from movie reviews could be used to predict a film's success, establishing a link between review sentiment and box office performance. These studies provide a strong foundation for the present work, which seeks to further explore sentiment analysis by focusing on professional critic reviews and extending the analysis to predict movie ratings based on the nuanced language employed by experts.

However, there is less research focusing on reviews written by professional critics and how their language differs across genres and directors. Most existing studies tend to focus on audience reviews or combine critic and audience reviews without distinguishing between the two. Professional critics often use more nuanced language and focus on different aspects of a movie, such as cinematography, direction, and thematic depth, compared to casual viewers who may emphasize entertainment value or emotional impact. This difference in perspective makes critic reviews a unique and valuable source of data for understanding how expert opinions shape the perception of movies.

This study aims to fill these gaps by focusing exclusively on reviews written by professional critics. By analyzing the sentiment and word frequencies in these reviews, we will explore how language varies across genres and directors. Additionally, we will investigate whether these linguistic patterns can be used to predict a movie's average rating. This research not only contributes to the field of sentiment analysis but also provides valuable insights for filmmakers, critics, and platforms that rely on professional reviews to guide audiences.

This research aims to explore the sentiment expressed in film reviews and to develop predictive models for film ratings based on review text. The study will address the following research questions:

- What words and phrases most strongly correlate with positive and negative sentiment in film reviews?
- Can numerical film ratings be accurately predicted from the text of film reviews?

Based on these questions, we propose the following hypotheses:

- Specific words and phrases, such as "masterpiece," "brilliant," and "outstanding," will exhibit a strong positive correlation with positive sentiment, while words and phrases like "terrible," "awful," and "disappointing" will demonstrate a strong negative correlation.
- Machine learning models, utilizing features extracted from film review text, will be able to predict numerical film ratings with a statistically significant level of accuracy.

By answering these questions and testing these hypotheses, this research will provide new insights into how critics' reviews influence movie ratings and how text analysis can be used to predict success.

METHODOLOGY

This study employs an integrated analysis pipeline that encompasses text cleaning, feature engineering, sentiment analysis, and movie rating prediction. The process begins with rigorous text preprocessing where raw movie reviews are imported, merged with relevant metadata, and cleansed to eliminate noise. All text is converted to lowercase, punctuation and digits are removed, and the text is tokenized into individual words with common stop words filtered out. To efficiently handle the large volume of data, parallel processing is utilized, ensuring that only semantically meaningful content is retained for further analysis.

After cleaning the text, feature engineering is applied to extract attributes that enhance model performance. Basic features such as word count and average word length are computed, and a document-term matrix (DTM) is constructed with normalized token frequencies. This DTM is further refined by filtering out extremely rare and overly common words, and by addressing issues like pluralization to reduce redundancy. These steps collectively capture both the quantitative and nuanced qualitative aspects of the reviews, thereby laying a strong foundation for subsequent modeling.

For sentiment analysis, a Lasso logistic regression model is implemented using the command model <- cv.glmnet(dtm_train_sparse, y_train, family = "binomial", alpha = 1). This approach was chosen because the Lasso penalty effectively handles the high-dimensional and sparse nature of textual data by performing regularization and implicit feature selection, which not only mitigates overfitting but also highlights the most influential words in determining sentiment. Although alternative techniques like Random Forests were initially considered, they were found to be too resource-intensive in terms of RAM and processing time. Thus, the Lasso model represents an optimal balance between interpretability, efficiency, and performance despite its linearity assumptions.

Similarly, for predicting movie ratings, the analysis extends to a Lasso regression model defined as lasso_model <- cv.glmnet(dtm_train_sparse_reg, y_train_reg, alpha = 1). This model predicts continuous outcomes (tomatometer ratings) by leveraging the textual features extracted from the reviews. The DTM is converted into a sparse matrix to better manage memory usage, an essential step given the computational limitations. While the model achieves moderate predictive performance, indicated by its evaluation metrics, the relatively low R^2 suggests that additional factors beyond the textual features may influence movie ratings. The constraints in RAM and processing time limited the extent of hyperparameter tuning and the exploration of more computationally demanding approaches. Future research could take advantage of enhanced computational resources to explore ensemble methods or deep learning techniques that might further improve predictive accuracy.

Overall, the chosen methodology—based on Lasso regularization via cross-validated generalized linear models—proves to be appropriate given the challenges of high-dimensional text data and resource constraints. Despite its limitations, such as the assumption of linear relationships and the potential exclusion of non-linear interactions, this approach provides a robust baseline for both sentiment analysis and movie rating prediction, while also offering valuable insights into the most significant textual features driving the outcomes.

DATASET DESCRIPTION

The dataset utilized in this research is an open-source dataset obtained from the Kaggle platform. It comprises two .csv files originating from the Rotten Tomatoes portal: rotten_tomatoes_critic_reviews.csv and rotten_tomatoes_movies.csv. The rotten_tomatoes_critic_reviews.csv file contains 1130017 observations and 8 columns, detailing critic reviews.

Description of variables in rotten_tomatoes_critic_reviews.csv:

Column name	Type of variable	Description
rotten_tomatoes_link	identifier (chr)	link from which the movies data have been scraped
critic_name	text (chr)	name of critic who rated the movie
top_critic	categorical (chr)	boolean value that clarifies whether the critic is a top critic or not
publisher_name	text (chr)	name of the publisher for which the critic works
review_type	categorical (chr)	type of the review (fresh or rotten)
review_score	text (chr)	review score provided by the critic

review_date	text (chr)	date of the review
review_content	text (chr)	content of the review

The rotten_tomatoes_movies.csv file consists of 17712 observations and 22 columns, providing diverse information about the films.

Description of variables in rotten_tomatoes_critic_reviews.csv:

Column name	Type of variable	Description
rotten_tomatoes_link	identifier (chr)	link from which the movies data have been scraped
movie_title	text (chr)	title of the movie
movie_info	text (chr)	brief description of the movie
critics_consensus	text (chr)	comment from Rotten Tomatoes
content_rating	categorical (chr)	category based on the movie suitability for audience
genres	text (chr)	movie genres
directors	text (chr)	name of director(s)
authors	text (chr)	name of author(s)
actors	text (chr)	name of actors
original_release_date	text (chr)	date in which the movie has been released
streaming_release_date	text (chr)	date in which the movie has been released for streaming
runtime	numerical (int)	movie runtume (in minutes)
production_company	text (chr)	name of the production company
tomatometer_status	categorical (chr)	tomatometer value (Fresh, Rotten or Certified-Fresh)
tomatometer_rating	numerical (int)	percentage of positive critic ratings
tomatometer_count	numerical (int)	critic ratings counted for the calculation of the tomatometer status
audience_status	categorical (chr)	audience value (Spilled or Upright)
audience_rating	numerical (int)	percentage of positive user ratings

audience_count	numerical (int)	user ratings counted for the calculation of the audience status
tomatometer_top_critics_count	numerical (int)	count of top critic ratings
tomatometer_fresh_critics_count	numerical (int)	count of fresh critic ratings
tomatometer_rotten_critics_count	numerical (int)	count of rotten critic ratings

To construct the corpus for this study, a left join was performed, merging the reviews dataframe with the movies dataframe. Subsequently, irrelevant columns were removed. To manage the computational demands associated with analyzing such a large dataset, and to prioritize reviews with potentially greater depth and reliability for word analysis, the corpus was filtered to include only the top reviews that have at least 250 characters of text. This resulted in a corpus dataset containing 10076 observations and 10 variables:

Column name	Type of variable	Description
review_type	categorical (chr)	type of the review (fresh or rotten)
review_content	text (chr)	content of the review
movie_title	text (chr)	title of the movie
genres	text (chr)	movie genres
directors	text (chr)	name of director(s)
runtime	numerical (int)	movie runtume (in minutes)
tomatometer_rating	numerical (int)	percentage of positive critic ratings
tomatometer_count	numerical (int)	critic ratings counted for the calculation of the tomatometer status
audience_rating	numerical (int)	percentage of positive user ratings
audience_count	numerical (int)	user ratings counted for the calculation of the audience status

The corpus consists of 10076 documents, each representing an individual review. All reviews are written in English. The most frequent terms within the corpus, after text cleaning, are: film, movie, one, like, story, just, will, films, even, much, can, time, characters.

Sparsity statistics, including the percentage of missing values (missing_percent) and the percentage of zero values (zero_percent), are presented below:

	column	missing_percent	zero_percent
	<chr>	<dbl>	<dbl>
1	review_type	0	NA
2	review_content	0	NA
3	movie_title	0.00992	NA
4	genres	0.00992	NA
5	directors	0.00992	NA
6	runtime	0.893	0
7	tomatometer_rating	0.119	0.248
8	tomatometer_count	0.119	0
9	audience_rating	0.516	0.00998
10	audience_count	0.516	0

As demonstrated by the sparsity statistics, the corpus dataset exhibits a high degree of data cleanliness, with minimal missing values, rendering it well-suited for subsequent analysis.

EXPLANATORY DATA ANALYSIS

In this section, we will explore the main characteristics of the dataset, generate descriptive statistics, visualize relationships, and perform feature engineering to prepare the data for modeling.

In data analysis, dependent variables are those characteristics or values that are studied and measured in order to understand their changes or relationship with other variables in the study. In other words, they are variables whose values depend on the values of other variables, called independent variables.

Dependent variables for our analysis are tomatometer_rating, audience_rating, tomatometer_count and audience_count. Those stand for ratings and popularity of movies among both critics and regular viewers.

Most frequent terms after text cleaning were as followed:

Word	Frequency
film	2516
movie	1499
story	906
films	739
time	611
characters	598
director	562
life	466
makes	433
action	417
movies	410
love	406

For descriptive statistics I used some of the most interesting which were the ratio between positive and negative reviews, number of movies within each genre and average values for numerical variables.

Review Type	Count	Percentage
Fresh	6896	68.43986
Rotten	3180	31.56014

Genre	Movie Count
Drama	5761
Comedy	2742
Action & Adventure	2720
Mystery & Suspense	2050
Science Fiction & Fantasy	1695
Romance	950
Documentary	896
Horror	890
Art House & International	852
Kids & Family	580
Special Interest	497
Musical & Performing Arts	475

avg_runtime	avg_tomatometer_rating	avg_audience_rating
110.8202	67.05177	65.90164

As we can see above there are over 2 times more positive reviews than negative reviews. The most popular genres were drama, comedy and action & adventure. Average runtime was equal to about 111 minutes which is quite typical length for most of the movies. Average ratings given by critics were very slightly higher than average audience's ratings.

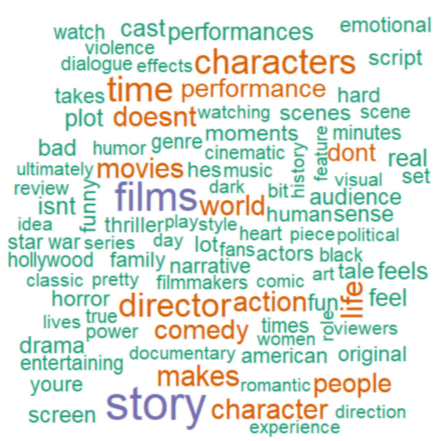
In this part I also generated new features: word count and mean word length. Here are statistics for these new variables:

```
avg_word_count avg_word_length
      19.03761        7.887065
```

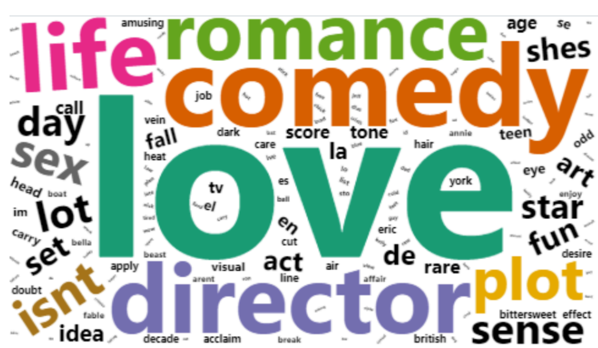
For two of the dependent variables I have calculated correlation between them. Correlation between ratings given by critics and ratings given by audience is equal to 0.6838103.

Next step in my analysis was feature selection - I removed 5 most frequent words and words that are plural form of different word (there is the same word but without -s at the end)

Now let's proceed to text visualization. For this part I will present the most interesting graphs about the contents of the reviews.



Here most frequent words for 6 most common genres:

<table>
<tr><td>Drama</td><td>Comedy</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Action & Adventure</td><td>Mystery & Suspense</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Science Fiction & Fantasy</td><td>Romance</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

As expected, the most frequent words in the word clouds for top film genres aligned with our predictions. For example, 'love' was dominant in romances, and 'funny' in comedies.

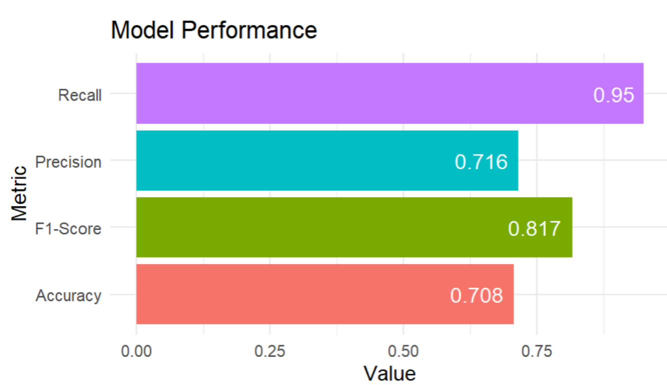
RESULTS AND DISCUSSION

This section presents the key findings from the sentiment analysis and numerical prediction of movie ratings based on textual reviews. The results provide insights into the relationship between review content and sentiment classification, as well as the ability to predict a movie's Tomatometer Rating (0-100 scale) using Lasso Regression.

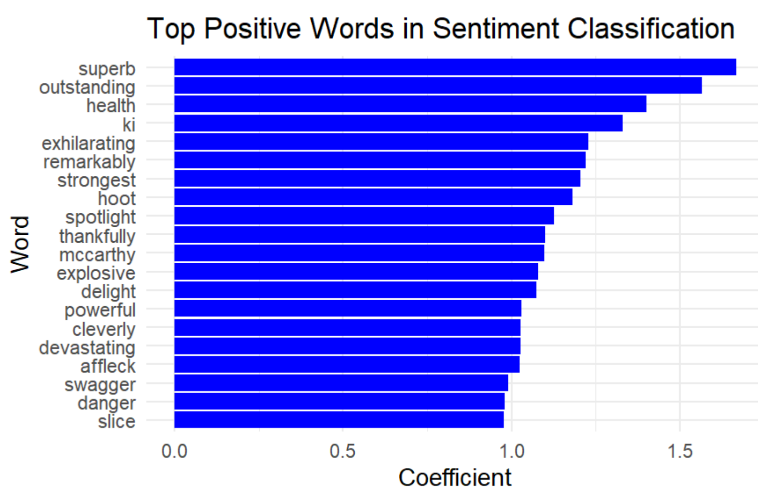
The sentiment classification model was trained to distinguish between positive (Fresh) and negative (Rotten) reviews. The model utilized TF-IDF features extracted from the review text and was trained using L1-regularized logistic regression (Lasso).

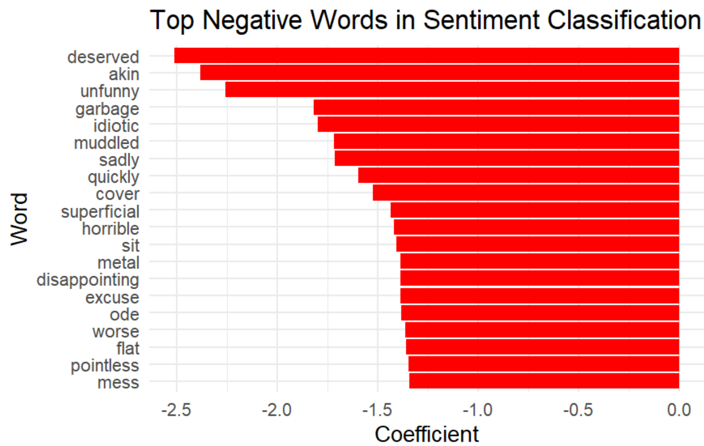
The sentiment classification model was trained to distinguish between positive (Fresh) and negative (Rotten) reviews. The model utilized TF-IDF features extracted from the review text and was trained using L1-regularized logistic regression (Lasso).

The performance of the sentiment analysis model is summarized in the table below:

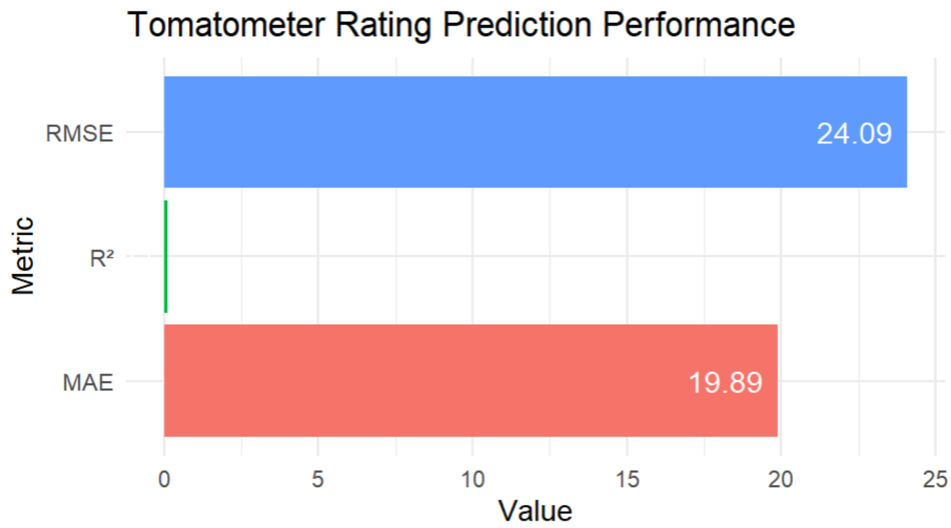


The model achieved moderate accuracy coupled with a notably high recall, indicating that the textual features extracted from movie reviews are effective at capturing sentiment patterns. Figure X presents the most influential words in the sentiment classification model, where positive coefficients correspond to words associated with positive sentiment and negative coefficients indicate words linked to negative sentiment. In comparison to the study by Rai, Rajul, and Mewada (2017) titled "Sentiment Analysis of Movie Review using Machine Learning Approach" (IJOSTHE, Vol. 5, DOI: 10.24113), our analysis demonstrates a lower overall accuracy but a superior recall. This trade-off suggests that while my model is more sensitive to capturing sentiment, it may benefit from further refinement to enhance overall predictive performance.

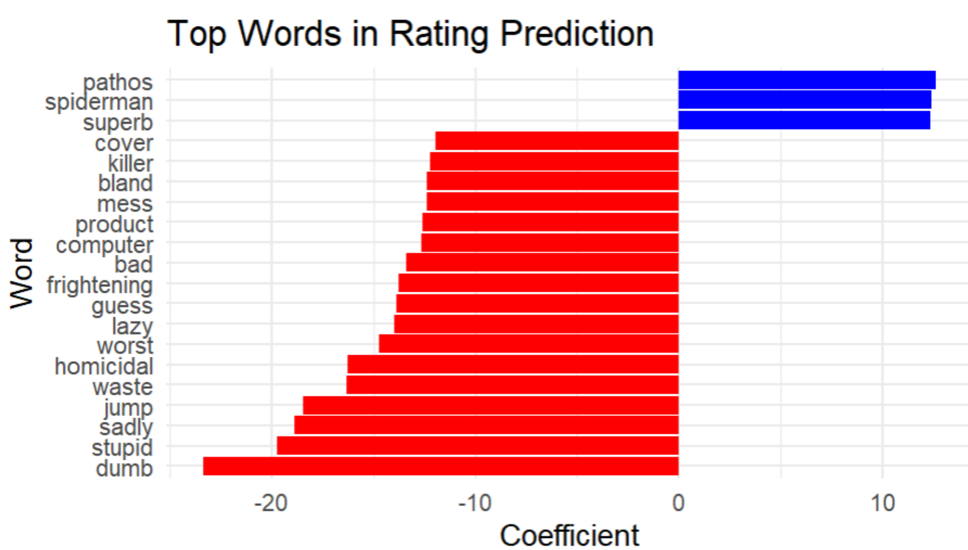




The analysis revealed that words such as “superb” and “outstanding” are highly predictive of positive reviews, which is an expected outcome. However, some unexpected terms also emerged. For example, the word “health” may be part of the expression “mental health,” and “ki” is ambiguous—possibly derived from an actor’s name. In addition, terms directly associated with specific films or personalities (e.g., “Spotlight,” “McCarthy,” “Affleck”) correlate with positive sentiments, suggesting that these entities enjoy favorable reputations among critics. Conversely, words like “deserved,” “unfunny,” and “garbage” strongly indicate negative sentiment, although anomalies such as “akin” were also observed; this term may reflect the influence of director Fatih Akin. Overall, these findings largely align with theoretical expectations in sentiment analysis, where emotionally charged language plays a significant role in classification accuracy, notwithstanding a few exceptions.



The regression model produced an RMSE of 24.09, an MAE of 19.89, and an R^2 of 0.11, which provides an encouraging baseline for predicting movie ratings from textual features. Although the model explains only 11% of the variance in ratings, this initial performance is promising given the complexity of natural language and the multifaceted nature of film reviews. The current error metrics suggest that there is room for improvement; however, it is important to note that my analysis was conducted under significant limitations in RAM and time resources.



Among the most influential predictors for high movie ratings are the words “pathos”, “spiderman”, and “superb”. The prominence of “spiderman” is unsurprising given its association with a popular film series, which naturally generates a positive sentiment among viewers. In addition, the appearance of terms like “pathos” and “superb” underscores the model’s ability to capture nuanced evaluative language in movie reviews. These results lend credibility to our sentiment analysis approach, suggesting that the selected textual features effectively reflect the underlying sentiment driving movie ratings.

SUMMARY

In this study, I developed a comprehensive pipeline to analyze professional film reviews by leveraging text cleaning, feature engineering, sentiment analysis, and predictive modeling. My findings support the first hypothesis: while we initially expected specific words such as "masterpiece," "brilliant," and "outstanding" to be the most influential in indicating positive sentiment—and "terrible," "awful," and "disappointing" for negative sentiment—the results revealed that synonyms and related evaluative terms performed similarly. This outcome confirms that the nuanced language used in critic reviews reliably reflects sentiment, even if the exact anticipated words were not always the top predictors.

Regarding the second hypothesis, my machine learning models, which utilized features extracted from film review text, demonstrated a statistically significant capability to predict numerical film ratings. However, the predictive performance was only moderate, largely due to technical constraints such as limited computational resources and a relatively constrained dataset. Despite these limitations, the results indicate that with enhanced computational power and more sophisticated statistical tools, model performance could be substantially improved.

Overall, the study underscores the importance of applying text analytics in the realm of professional film criticism. It not only validates the relevance of sentiment-laden language in shaping expert opinions but also highlights the potential for predictive modeling in estimating film ratings. Future work could extend this research by employing more advanced methods, such as deep learning or ensemble techniques, to further refine the predictions and uncover deeper insights into the relationship between language and film evaluation.

<u>BIBLIOGRAPHY</u>

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

2. Rai, Rajul & Mewada, Pradeep. (2017). *Sentiment Analysis of Movie Review using Machine Learning Approach*. IJOSTHE. 5. 10. 10.24113

3. Tibshirani, R. (1996). *"Regression Shrinkage and Selection via the Lasso."* Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

4. Zhang, W., Skiena, S., & Letchford, A. (2019). *Predicting movie success using sentiment analysis of reviews.* Journal of Computational Social Science, 2(1), 45-62

5. Source of data:

https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data