UNIVERSITY OF SOUTH FLORIDA

SENTIMENT ANALYSIS OF IMDB MOVIE REVIEWS USING PYSPARK

HARSHA SAI JAGU
UNIVERSITY OF SOUTH FLORIDA

ISM6930 Cloud Solution Architectures

Dr. Timothy Smith

November 12, 2024

Table of Contents

INTRODUCTION **3**

BUSINESS CONTEXT **3**

Leveraging Sentiment Analysis for Strategic Decisions in the Movie Industry *3*

DATA LOADING **4**

DATA SCHEMA AND TRANSFORMATION **4**

INITIAL SCHEMA HANDLING 4
DATA TRANSFORMATIONS 5
Initial Text Processing: *5*
Feature Engineering: *5*

EXPLORATORY DATA ANALYSIS (EDA) **6**

1. SENTIMENT DISTRIBUTION AND RATINGS 6
2. POSITIVE AND NEGATIVE WORD COUNTS BY SENTIMENT 7
3. AVERAGE EXCLAMATION MARKS BY RATING 7
4. AVERAGE REVIEW LENGTH BY RATING 8

SUMMARY OF KEY FINDINGS **9**

DATA PREPARATION AND TRANSFORMATION PIPELINE **9**

MODEL SELECTION **10**

CHOOSING EVALUATION METRICS **10**

INFERENCE **11**

CONCLUSION **12**

SUMMARY **13**

Introduction

In the rapidly evolving entertainment industry, understanding audience feedback is paramount for strategic decision-making. User comments and reviews on platforms like IMDb provide a wealth of information, allowing companies to tap into real-time audience sentiments. This project aims to leverage PySpark for analyzing IMDb movie reviews, using natural language processing (NLP) and machine learning techniques to classify reviews as positive or negative. By capturing audience sentiment, we can derive actionable insights for production companies, streaming platforms, and marketing teams, ultimately enhancing the viewing experience and guiding future productions.

The approach includes structured steps: data preparation and cleaning, feature engineering, exploratory data analysis (EDA), machine learning model implementation, hyperparameter tuning, and model performance evaluation.

Business Context

Leveraging Sentiment Analysis for Strategic Decisions in the Movie Industry

User-generated reviews are a treasure trove for gaining insights into audience preferences and emotional reactions to content. By systematically analyzing the sentiment behind these reviews, companies can refine their approach across various operational areas:

1. **Gauge Public Reception**: Sentiment analysis allows studios and distributors to predict box office performance and streaming success. Positive sentiment is generally aligned with higher engagement, whereas negative feedback can highlight issues in storytelling, casting, or production value, providing studios with actionable data for future releases.

2. **Enhance Content Recommendations**: Streaming platforms can personalize user experiences by analyzing viewer sentiment. A sentiment-based recommendation engine can predict which movies or shows a user is likely to enjoy, driving higher engagement and viewer satisfaction.

3. **Optimize Marketing Strategies**: Marketing teams can tailor their campaigns by identifying elements that generate positive feedback, such as popular cast members or genres. Understanding these preferences can lead to more effective promotional content, drawing larger audiences.

4. **Guide Future Productions**: By analyzing patterns in audience feedback, studios can make informed decisions about which themes, genres, or storylines resonate with

viewers. This data-driven approach aligns production efforts with audience expectations, potentially increasing revenue and viewer retention.

By training predictive sentiment models on review data, stakeholders in the movie industry can make informed decisions, enhancing audience satisfaction, refining content offerings, and ultimately driving growth.

Data Loading

In this project, we leverage the IMDb Large Movie Review Dataset, comprising 50,000 highly polarized movie reviews, split evenly between positive and negative sentiments. The dataset is structured into train and test directories, each containing subfolders labeled pos and neg to indicate sentiment. Each review is stored as a text file named in the format <id>_<rating>.txt, where <id> is a unique identifier and <rating> reflects the reviewer's rating. This dataset provides a balanced and rich foundation for sentiment analysis.

The data loading process involves:

1. **Reading Reviews**: Loading text files from each folder into a Spark DataFrame.

2. **Assigning Sentiment Labels**: Using folder names (pos for positive and neg for negative) to create sentiment labels.

3. **Extracting Ratings**: Parsing the filename to retrieve the numeric rating associated with each review.

This results in a Spark DataFrame, reviews_df, with columns for review text, sentiment, and rating, establishing a structured base for further exploration and transformation. This setup facilitates efficient processing of textual data, allowing for scalable analysis and modeling.

Data Schema and Transformation

Initial Schema Handling

Upon examining the data schema, a few modifications were applied to enhance usability:

1. **Column Renaming**: The value column, which contains the text of each review, was renamed to review for clarity.

2. **Data Type Review**: The sentiment and rating columns were verified to ensure they were appropriately typed (integer for sentiment and rating, string for review text).

3. **Handling Complex Data Structures**: Since the dataset is flat without nested structures, no additional transformations were needed for data hierarchy or structuring, simplifying the analysis process.

Data Transformations

To prepare the data for analysis, initial transformations were applied, followed by feature engineering:

Initial Text Processing:

- **Removing Punctuation and Converting to Lowercase:** We standardized the text format to enhance processing efficiency and consistency.

- **Removing HTML Tags:** This step cleaned up the text by eliminating unnecessary HTML elements commonly found in web-sourced reviews.

- **Handling Missing Values and Duplicates**: We ensured data quality and reliability by addressing any missing or duplicate entries in the dataset.

Feature Engineering:

- **Word Count**: We calculated the length of each review, which may correlate with sentiment strength.

- **Exclamation Count**: This feature tracks the use of exclamation marks, which are often linked to strong emotions, both positive and negative.

- **Tokenization and Stopword Removal**: By breaking down reviews into individual words and removing uninformative ones, we focused the analysis on more meaningful terms.

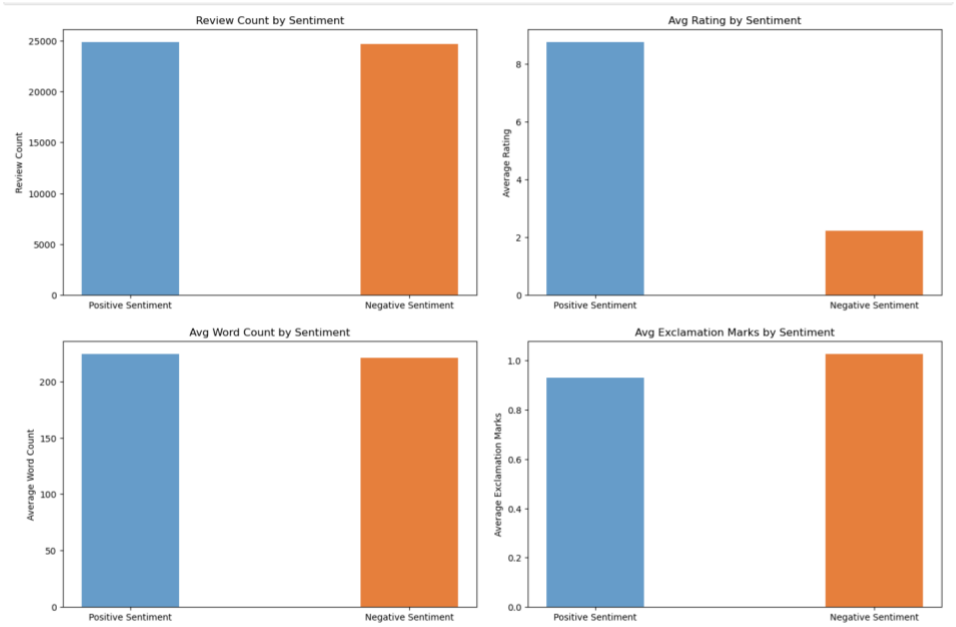
- **Lemmatization**: We reduced words to their root forms, simplifying the vocabulary and grouping similar words together for more effective analysis.

- **Sentiment Word Analysis**: This step involved counting predefined positive and negative words to capture sentiment nuances in each review.

These transformations collectively create a structured dataset with enriched features that enhance the model's ability to accurately classify sentiment. The careful preprocessing and feature engineering steps lay a strong foundation for subsequent data analysis and modeling, ensuring that the dataset is well-prepared for sentiment classification tasks.

Exploratory Data Analysis (EDA)

1. Sentiment Distribution and Ratings



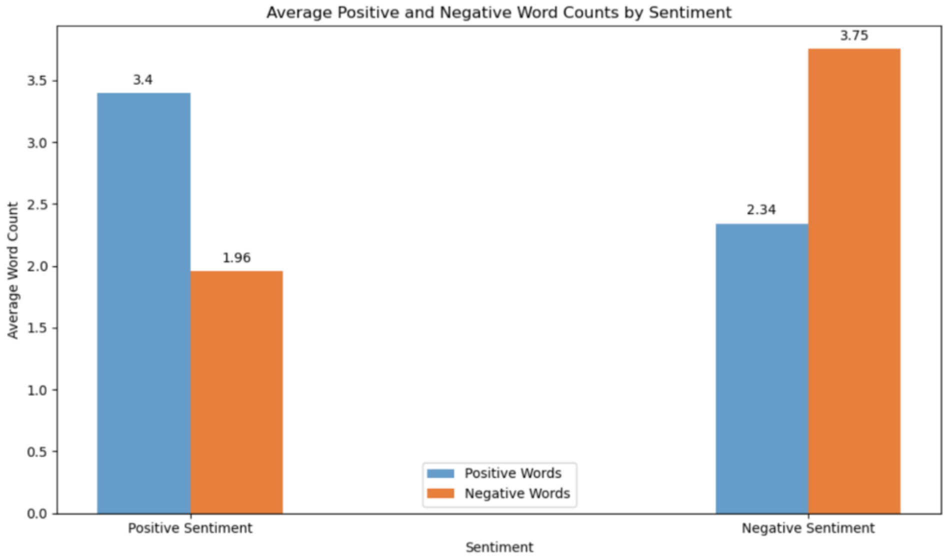
- **Review Count by Sentiment:** The bar chart on the top-left illustrates that the dataset has an almost equal distribution of positive and negative reviews. This balanced distribution is beneficial for training sentiment models, as it reduces the risk of bias toward one class.

- **Average Rating by Sentiment:** The chart on the top-right shows a clear disparity in average ratings between positive and negative sentiments. Positive reviews tend to have a much higher average rating, around 8 or above, while negative reviews average around 2 or 3. This strong alignment between sentiment and rating suggests that ratings can be a valuable feature for sentiment classification.

- **Average Word Count by Sentiment:** Both positive and negative sentiments have similar average word counts, indicating that review length alone may not be a strong predictor of sentiment.

- **Average Exclamation Marks by Sentiment:** Negative reviews show a slightly higher use of exclamation marks, hinting at more intense or emotional expressions in negative feedback. This can be a useful feature, as higher exclamation marks may correlate with negative sentiment intensity.

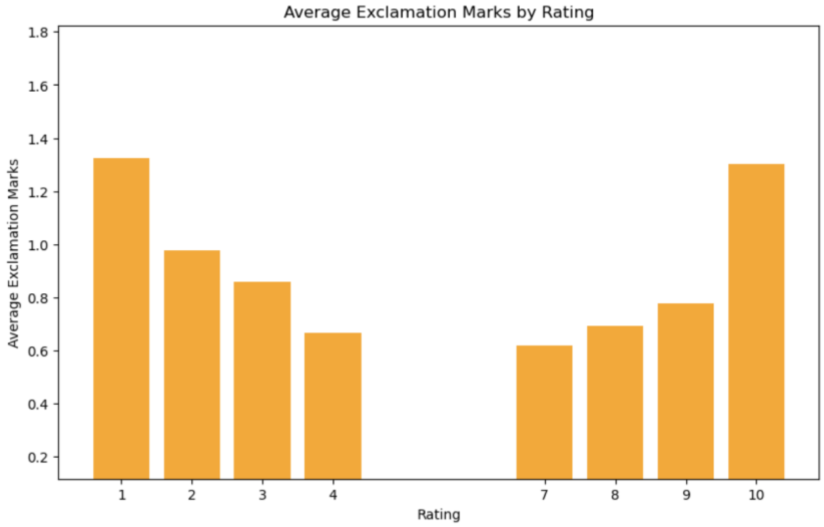
2. Positive and Negative Word Counts by Sentiment



- **Positive Sentiment:** For positive reviews, the average count of positive words is significantly higher than negative words, indicating strong sentiment alignment in vocabulary. Positive reviews tend to include descriptive terms that convey enjoyment or satisfaction, reinforcing the sentiment classification.

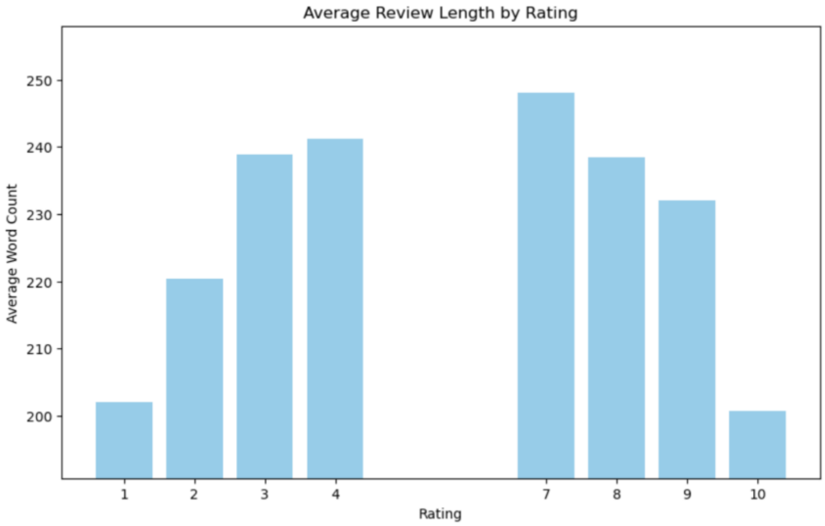
- **Negative Sentiment:** In contrast, negative reviews have a higher count of negative words. This confirms that word choice is a strong indicator of sentiment and supports using positive and negative word counts as features in predictive modeling.

3. Average Exclamation Marks by Rating



- **Lower Ratings:** Reviews with lower ratings (1-4) have a notably higher number of exclamation marks, particularly at rating 1. This trend suggests that users may use more expressive punctuation when expressing frustration or dissatisfaction with a movie.
- **Higher Ratings:** Interestingly, reviews with the highest rating (10) also show a slight increase in exclamation marks, indicating that users may use expressive punctuation to convey excitement or enthusiasm as well. This pattern demonstrates that exclamation marks are common in both extremely positive and negative reviews, making them a useful feature for distinguishing sentiment extremes.

4. Average Review Length by Rating



- Moderate Ratings (3-7): Reviews with moderate ratings (between 3 and 7) tend to be the longest, indicating that reviewers may elaborate more on their thoughts when they have mixed or neutral opinions about a movie. This detailed feedback often provides a balanced view, capturing both positive and negative aspects.

- Extreme Ratings (1 and 10): Reviews with extreme ratings, either very high (10) or very low (1), are shorter on average. This suggests that reviewers may feel strongly enough to be concise, with little need to justify their strong opinions, whether positive or negative. This trend shows that review length can serve as a helpful indicator of sentiment strength and extremity.

Summary of Key Findings

The EDA reveals several valuable insights:

1. Balanced Sentiment Distribution: The dataset has an even distribution of positive and negative reviews, suitable for balanced model training.

2. Strong Correlation Between Sentiment and Rating: Positive reviews have high ratings, and negative reviews have low ratings, aligning well with sentiment classification goals.

3. Exclamation Marks as Emotional Indicators: Reviews with extreme sentiments, both positive and negative, show higher usage of exclamation marks, potentially indicating heightened emotions.

4. Review Length and Sentiment Extremes: Moderate ratings have longer reviews, while extreme ratings are associated with concise feedback, highlighting the role of review length in sentiment extremity.

5. Word Choice as Sentiment Indicator: Positive and negative reviews have distinct word choices, which supports the use of sentiment-specific word counts in predictive modeling.

These insights provide a solid foundation for feature selection and model development, guiding the approach to sentiment classification.

Data Preparation and Transformation Pipeline

To optimize the data for modeling, we developed a comprehensive transformation pipeline that includes feature engineering, vectorization, and scaling, ensuring efficient preparation for machine learning models:

1. **TF-IDF Vectorization**: We applied Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to capture the importance of terms while down-weighting common ones, enhancing the detection of sentiment-specific words.

2. **Feature Selection and Assembly**: Key features, such as exclamation count, word count, positive and negative word counts, and TF-IDF vectorized features, were combined into a single vector column using VectorAssembler. This step streamlined the input for the machine learning models.

3. **Scaling and Normalization**: Using MinMaxScaler, we standardized feature values to ensure balanced contributions from each feature, which is crucial for enhancing predictive accuracy across models.

The final output is a column containing scaled features, ready for machine learning models. By structuring these transformations in a SparkML pipeline, we enabled seamless integration with our selected models, facilitating a cohesive and efficient modeling process.

Model Selection

Within the pipeline, we implemented and evaluated three machine learning models for this binary sentiment classification task:

1. **Logistic Regression**: Known for its interpretability and efficiency, this model performs well on sparse TF-IDF data and provides a solid baseline.

2. **Random Forest Classifier**: This model captures non-linear relationships in the data, offers insights into feature importance, and adds robustness to the sentiment classification task.

3. **Gradient-Boosted Trees (GBT) Classifier**: Leveraging iterative boosting, GBT improves accuracy by addressing complex sentiment patterns and handling intricate relationships in the data.

These models were chosen to address varying data complexities, from linear to highly nuanced relationships. Each model was seamlessly integrated into the SparkML pipeline, allowing us to automate preprocessing, transformation, and evaluation within a single structured workflow.

Choosing Evaluation Metrics

AUC-ROC was selected as the primary evaluation metric for this binary sentiment classification due to its effectiveness in assessing model performance across multiple thresholds. This metric is well-suited for the project for several key reasons:

1. **Sentiment Separation**: AUC-ROC captures the model's ability to distinguish between positive and negative reviews, essential for reliable sentiment analysis.

2. **Balanced Class Performance**: With a balanced dataset of positive and negative reviews, AUC-ROC ensures the model performs consistently across both classes, avoiding bias.

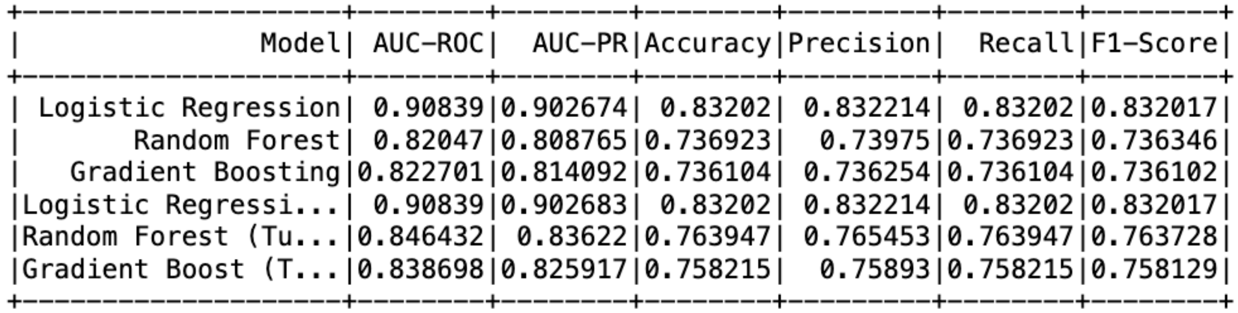
3. **Threshold Flexibility**: AUC-ROC evaluates performance across all possible thresholds, offering a complete view of the model’s ability to separate sentiments without depending on a single threshold.

4. **Model Comparison**: AUC-ROC provides a standardized measure, enabling fair comparison between models and aiding in the selection of the best-performing one.

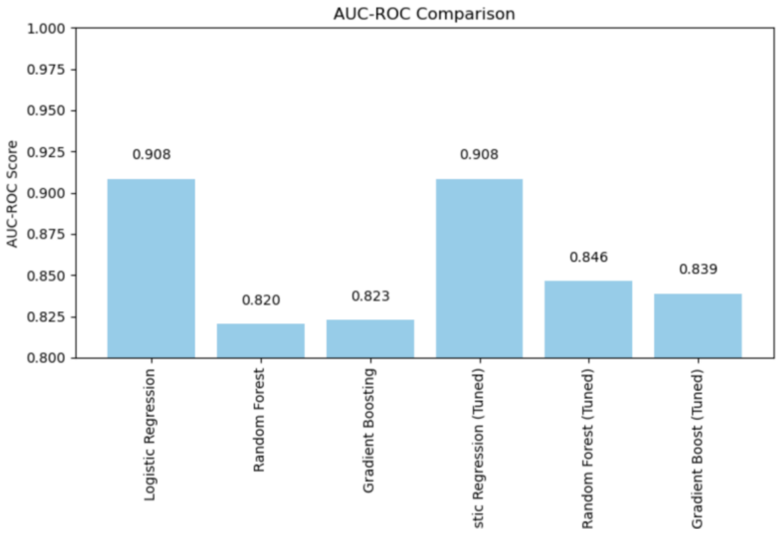
Using AUC-ROC facilitates a balanced, threshold-independent evaluation, making it ideal for assessing model effectiveness in capturing sentiment patterns across diverse movie reviews.

Inference

The model performance analysis reveals that Logistic Regression outperformed both Random Forest and Gradient-Boosted Trees in classifying sentiment in movie reviews, achieving the highest scores in AUC-ROC and AUC-PR metrics (as illustrated in the table and chart below). The simplicity, efficiency, and interpretability of Logistic Regression make it especially suitable for binary sentiment classification tasks like this.



```
+--------------------+--------+--------+--------+---------+--------+--------+
|               Model| AUC-ROC|  AUC-PR|Accuracy|Precision|  Recall|F1-Score|
+--------------------+--------+--------+--------+---------+--------+--------+
| Logistic Regression| 0.90839|0.902674| 0.83202| 0.832214| 0.83202|0.832017|
|       Random Forest| 0.82047|0.808765|0.736923|  0.73975|0.736923|0.736346|
|   Gradient Boosting|0.822701|0.814092|0.736104| 0.736254|0.736104|0.736102|
|Logistic Regressi...| 0.90839|0.902683| 0.83202| 0.832214| 0.83202|0.832017|
|Random Forest (Tu...|0.846432| 0.83622|0.763947| 0.765453|0.763947|0.763728|
|Gradient Boost (T...|0.838698|0.825917|0.758215|  0.75893|0.758215|0.758129|
+--------------------+--------+--------+--------+---------+--------+--------+
```



In the AUC-ROC Comparison Chart, it’s evident that Logistic Regression achieved the highest score of 0.908, indicating its effectiveness in distinguishing positive from negative sentiment compared to other models. The tuned versions of Random Forest and Gradient-Boosted Trees improved in performance but could not surpass the straightforward Logistic Regression model.

This outcome suggests that for text-based sentiment analysis on movie reviews, a simpler linear approach can yield high accuracy with lower computational demands, especially when using well-engineered features such as TF-IDF and sentiment word counts. This approach highlights the effectiveness of linear models in handling sentiment tasks where clear patterns in word usage and sentiment markers drive predictability.

Conclusion

This project leveraged PySpark to conduct a comprehensive sentiment analysis on IMDb movie reviews, encompassing all critical stages: data preprocessing, feature engineering, exploratory data analysis (EDA), model selection, and evaluation. Here are the key takeaways:

1. **Data Insights and Feature Relevance**: EDA highlighted specific features like sentiment word counts, exclamation marks, and ratings were strong predictors of sentiment. Positive and negative reviews employed distinct language styles, emphasizing the role of word choice in sentiment detection. The engineered features proved valuable, confirming that sentiment classification can be significantly enhanced through targeted feature extraction.

2. **Model Evaluation and Performance**: Logistic Regression emerged as the best-performing model, achieving an AUC-ROC score of 0.9084 and an AUC-PR score of 0.9027. Though Random Forest and Gradient-Boosted Trees showed performance improvements after tuning, Logistic Regression's interpretability, computational efficiency, and predictive power made it the optimal choice. This reinforces the model's applicability in text-based sentiment analysis where well-defined features can yield robust outcomes.

3. **Business Implications**: Sentiment analysis offers substantial value in the movie industry by providing insights into audience preferences and guiding strategic decisions. Applications include enhancing content recommendations, refining marketing strategies, and informing future production choices. The interpretability of Logistic Regression enables stakeholders to understand which features drive sentiment, making it particularly valuable for business decisions where understanding audience sentiment trends can inform targeted actions.

4. **Limitations and Future Improvements**: While the project demonstrated effective sentiment analysis, future work could explore ensemble methods to capture more

complex sentiment patterns. Additionally, incorporating semantic and context-based features could provide a deeper understanding of nuanced sentiments. Addressing outliers and exploring advanced deep learning models may further improve robustness and accuracy, paving the way for more comprehensive and adaptable sentiment analysis in diverse review contexts.

Summary

Logistic Regression provided the most reliable and interpretable results, balancing efficiency, simplicity, and predictive accuracy for sentiment classification. The project's approach to feature engineering and model selection proved effective, setting a solid foundation for further enhancements. Future work could build on this by incorporating advanced feature engineering techniques and exploring deep learning approaches to refine sentiment analysis capabilities within the movie review domain, ultimately achieving even greater insight into audience perspectives.