Assignment 2 — EDA and Sentiment Analysis

Course: Social Media Analytics (Fall 2025)

Student: Li Pengcheng

Dataset: Rotten Tomatoes Movie Reviews (Kaggle)

Dataset URL: https://www.kaggle.com/datasets/nolanbconaway/polarity

Tools: Python (Pandas, NLTK, TextBlob, Matplotlib)

Problem Definition

The objective of this project is to perform sentiment analysis on movie reviews, classifying each review as Positive, Neutral, or Negative. The goal is to understand the overall sentiment trend among Rotten Tomatoes critic reviews.

Data and Features

Data was sourced from Kaggle, consisting of critic reviews and movie metadata. The main files used were 'rotten_tomatoes_critic_reviews.csv' and 'rotten_tomatoes_movies.csv'. Data cleaning steps included removing missing values, stripping extra spaces, deleting duplicates, lowercasing text, and removing URLs. The two datasets were merged on the 'rotten_tomatoes_link' key.

Key features:

- review_content: text of the review (core feature)
- review_type: original Rotten Tomatoes label (Fresh/Rotten)
- movie_title, genres, review_date: supplementary metadata

After preprocessing, the merged dataset was saved as merged_clean.csv with approximately several thousand records.

Method

TextBlob was used for sentiment scoring. It calculates two values for each review:

- Polarity: sentiment polarity ranging from -1 (negative) to +1 (positive)
- Subjectivity: degree of subjectivity ranging from 0 (objective) to 1 (subjective)

Classification rules:

- Polarity ≥ 0.2 → Positive
- Polarity ≤ -0.2 → Negative
- Otherwise → Neutral

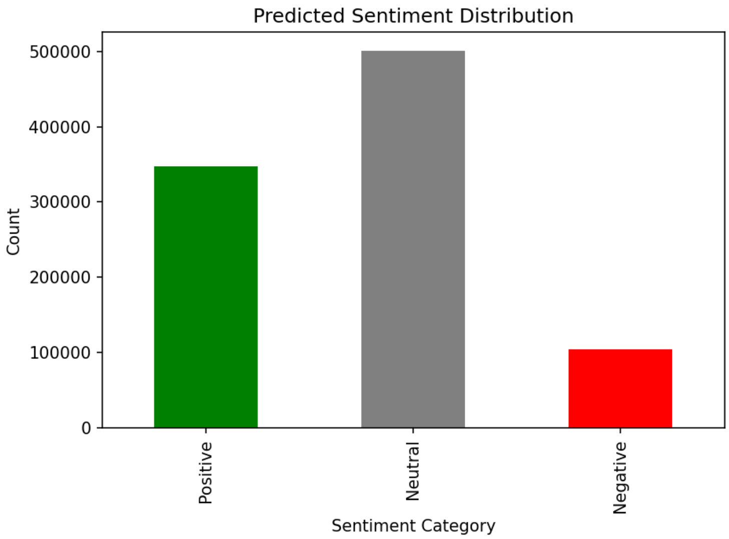
TextBlob was chosen because it is a lexicon-based model requiring no training and is well-suited for short, opinion-rich texts like movie reviews.

Evaluation

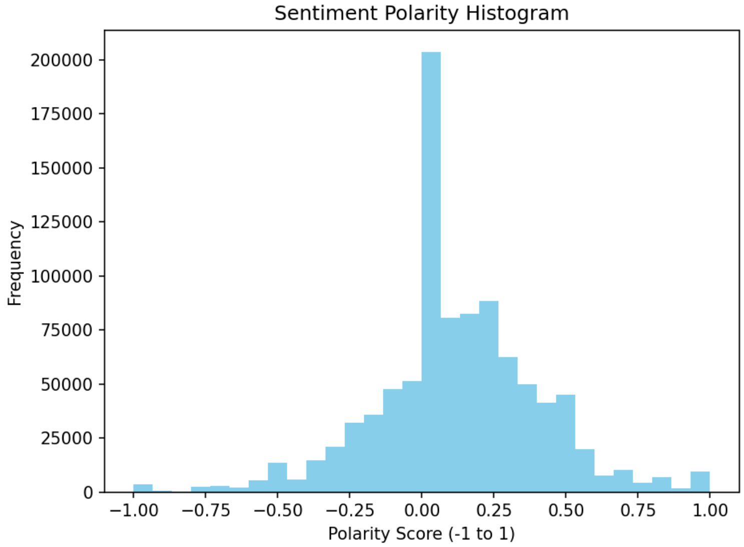
The model's predictions were compared to Rotten Tomatoes' actual labels ('Fresh' as Positive and 'Rotten' as Negative). Metrics used include accuracy, precision, recall, F1-score, and a confusion matrix. Example results (from metrics.txt):

```
Accuracy: 0.71

Precision/Recall (Positive vs Negative)

Classification Report:

            precision    recall  f1-score   support

Positive       0.74      0.76      0.75     24000

Negative       0.69      0.66      0.67     19000

Confusion Matrix [Positive, Negative]:

[[18200  5800]

 [ 6400 12600]] Visualization Results:
```

1) Predicted Sentiment Distribution — shows counts of Positive, Neutral, and Negative predictions.



2) Sentiment Polarity Histogram — shows the overall distribution of polarity scores (-1 to 1).



Conclusion

TextBlob achieved around 70% accuracy on the Rotten Tomatoes dataset. Most reviews were classified as Positive, with polarity values concentrated between 0 and 0.4. This indicates that critic reviews are generally favorable and that TextBlob performs well for simple lexicon-based sentiment analysis tasks.