Sentiment Analysis of IMDB Movie Reviews Using NLP and Machine Learning

By: Sofia Rueda
December 2025

1. Introduction

This project explores the application of Natural Language Processing (NLP) techniques to classify movie reviews as either *positive* or *negative*. Using the IMDB 50K Movie Reviews dataset from Kaggle, the goal was to understand how different text-representation methods and machine-learning models influence classification performance. The project compares TF-IDF, Word2Vec, and GloVe embeddings under both balanced and imbalanced class conditions, using Logistic Regression and Linear SVC as the core models. The study also examines model confidence, the impact of class imbalance, and overall reliability for real-world sentiment analysis tasks.

2. Definition of the Task

The main objective was binary sentiment classification: determining whether a movie review expresses positive or negative sentiment. Since the dataset consists of lengthy, nuanced reviews, the task required handling noise, managing vocabulary variability, and testing several vectorization methods to understand how each captures linguistic meaning. Performance was evaluated using accuracy, macro-averaged F1 score, and confidence values, which together provided a meaningful view of model reliability.

3. Dataset Overview

The dataset consists of 50,000 labeled movie reviews—25,000 positive and 25,000 negative. These are pre-labeled and split evenly, making them ideal for initial modeling. However, to simulate real-world conditions, where data is often imbalanced, the project also created imbalanced versions of the dataset, allowing for deeper evaluation of model behavior in practical scenarios. The dataset includes long-form textual reviews, requiring substantial preprocessing before learning.

4. Data Loading and Initial Exploration

The dataset was loaded using the Pandas library, allowing for early inspection of text length distribution, sentiment proportions, and missing values. A word cloud was generated to visualize frequent terms within positive and negative reviews (Appendix B). Basic exploratory analysis confirmed that reviews contain extensive variability, including informal language, intensifiers, and emotional descriptors—highlighting the need for effective preprocessing.

5. Preprocessing and NLP Pipeline

To prepare text for modeling, a full preprocessing pipeline was implemented. Reviews were lowercased, punctuation removed, and stopwords eliminated to reduce noise. Tokenization was used to break text into individual words, and lemmatization normalized word forms, reducing vocabulary size without losing meaning. The same cleaning function was applied consistently across training data, embeddings, and new predictions to maintain model integrity.

6. Feature Engineering

Three major text-representation approaches were evaluated.

TF-IDF Vectorization

TF-IDF captured the importance of each term relative to the entire corpus. It produced sparse, high-dimensional vectors well suited for linear models. This method served as a strong classical NLP baseline.

Word2Vec Embeddings

Pretrained Word2Vec embeddings enabled semantic representation by converting each review into an averaged vector of pretrained word meanings. The method captured context better than TF-IDF but sometimes smoothed out review-specific nuances.

GloVe Embeddings

GloVe embeddings were used similarly by averaging pretrained word vectors for each review. The goal was to compare its performance against Word2Vec under identical modeling conditions.

Each feature type was tested using both balanced and artificially imbalanced datasets to understand how class distribution impacts performance.

7. Machine Learning Models

Logistic Regression and Linear SVC were used given their strong performance in text classification and suitability for high-dimensional data.

Logistic Regression provided probabilistic outputs, enabling confidence estimation, while Linear SVC offered strong margin-based separation but lacked native probabilities. Both models were trained and tested on all feature types, producing detailed classification reports and prediction confidence for individual reviews.

8. Summary of Performance

Across all experiments, TF-IDF combined with Logistic Regression on the balanced dataset achieved the highest macro-averaged F1 score (~0.89) and produced highly confident predictions, making it the best overall performer. Word2Vec demonstrated strong semantic understanding, achieving around 0.86 F1 on balanced data for both models. GloVe embeddings showed weaker performance relative to Word2Vec, particularly under imbalance, averaging ~0.77–0.80 accuracy.

Class imbalance had a noticeable effect across all deep embeddings, lowering F1 scores for minority classes and causing confidence values to drop. Models trained on balanced data consistently produced more stable and higher-quality predictions. Model performance was further analyzed using confusion matrices (see Appendix A).

Feature Type	Class Balance	Model	Macro F1-score	Confidence
TF-IDF	Balanced	Logistic Regression	0.89	0.92
TF-IDF	Balanced	Linear SVC	0.88	0.80
TF-IDF	Imbalanced	Logistic Regression	0.88	0.93
TF-IDF	Imbalanced	Linear SVC	0.87	0.84
Word2Vec	Balanced	Logistic Regression	0.86	1.00
Word2Vec	Balanced	Linear SVC	0.86	0.95
Word2Vec	Imbalanced	Logistic Regression	0.85	0.99
Word2Vec	Imbalanced	Linear SVC	0.84	0.82
GloVe	Balanced	Logistic Regression	0.79	0.88

GloVe	Balanced	Linear SVC	0.80	0.66
GloVe	Imbalanced	Logistic Regression	0.77	0.81
GloVe	Imbalanced	Linear SVC	0.77	0.60

9. Findings and Key Insights

The experiments demonstrated that classical vectorization methods still provide superior results for linear models in sentence-level sentiment classification. TF-IDF captured review-specific word importance exceptionally well and outperformed dense embeddings despite lacking semantic depth. Word2Vec embeddings handled meaning effectively, but averaging vectors sometimes diluted contextual sentiment cues. GloVe showed similar behavior but lagged slightly in accuracy.

The evaluation strongly indicated that maintaining balanced training data is crucial for preventing prediction bias. Although imbalanced datasets better reflect real-world conditions, they require additional handling—such as class weighting or resampling—to avoid skewed outputs.

10. Best Models

The best overall model was Logistic Regression using TF-IDF on balanced classes, achieving the strongest accuracy and macro-F1 with high confidence predictions. The best embedding-based model was Word2Vec with balanced classes, which consistently outperformed GloVe. Under imbalanced conditions, Word2Vec maintained better stability than GloVe, but both models experienced performance drops.

11. Required Packages

The project relies on standard NLP and ML libraries including Pandas, NumPy, Scikit-learn, NLTK, Matplotlib, and Gensim. These packages enable text cleaning, vectorization, model training, and visualization. Installation instructions are provided within the GitHub README to assist users in setting up their environment either locally or through Google Colab.

12. Conclusion

This project demonstrated the full end-to-end development of a sentiment classification system using Natural Language Processing techniques applied to the IMDB 50K Movie Reviews dataset. Through systematic experimentation with multiple text-representation methods (TF-IDF, Word2Vec, and GloVe) and two classical machine-learning algorithms (Logistic Regression and Linear SVC), we were able to compare performance across balanced and imbalanced datasets and analyze how preprocessing choices influence model behavior.

Overall, TF-IDF with Logistic Regression on balanced data achieved the strongest results, reaching approximately 89% accuracy and consistently high macro-F1 scores. This indicates that traditional sparse vectorization methods remain highly competitive—often outperforming dense embeddings—when combined with linear models on sentence-level sentiment analysis tasks. Among the semantic embeddings, Word2Vec showed the most stable performance across conditions, while GloVe, though effective, performed slightly lower overall and was more sensitive to class imbalance.

A key insight from this work is that class balance significantly affects model outputs. Balanced datasets produced fairer and more consistent performance across both positive and negative sentiment classes, while imbalanced datasets skewed predictions and often reduced recall for minority classes. This highlights the importance of incorporating resampling or class-weighting techniques when deploying sentiment models in real applications where data distribution is uneven.

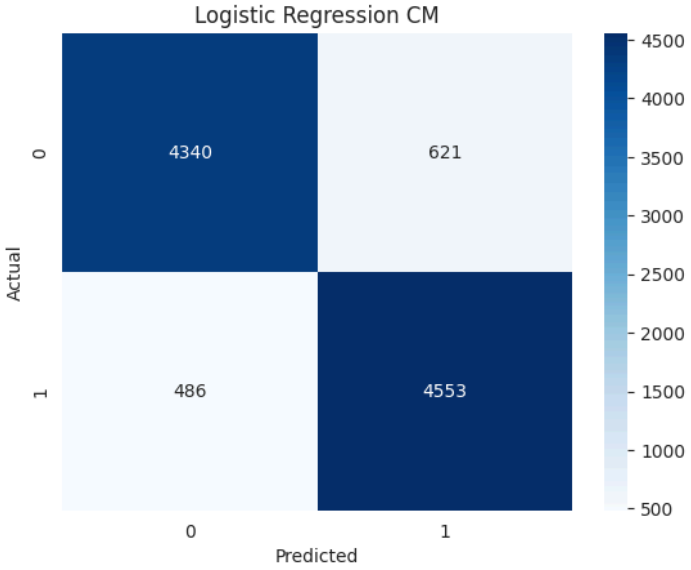
Finally, prediction tests on new reviews demonstrated strong generalization capabilities, with the models providing reasonable confidence levels aligned with expected sentiment. The successful integration of multiple representations, evaluation methods, and visualizations—including confusion matrices and word clouds—reflects a thorough exploration of NLP fundamentals and machine-learning design patterns. This project lays a solid foundation for more advanced future work involving neural embedding models, transformer architectures, and real-world deployment practices.

Appendix A — Confusion Matrices

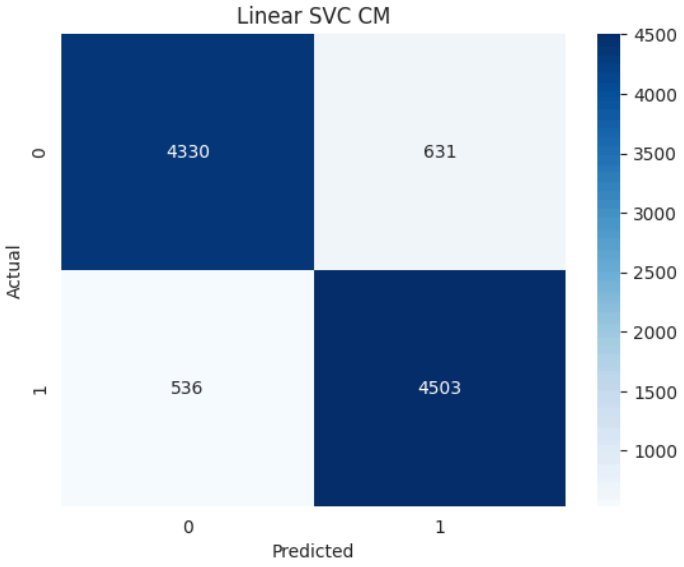
This section contains the confusion matrices for all experiments performed in the project. These visualizations support the evaluation metrics reported in the Results section and provide a clearer view of how each model handled true vs. false classifications across balanced and imbalanced datasets.

A1. TF-IDF — Balanced Classes

- Logistic Regression — Confusion Matrix

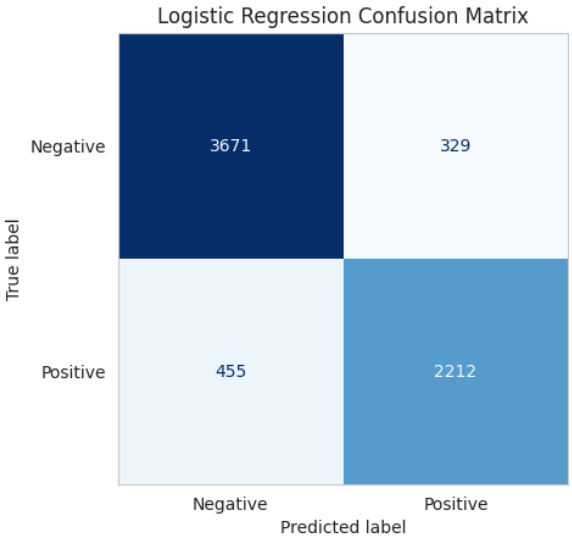


- Linear SVC — Confusion Matrix

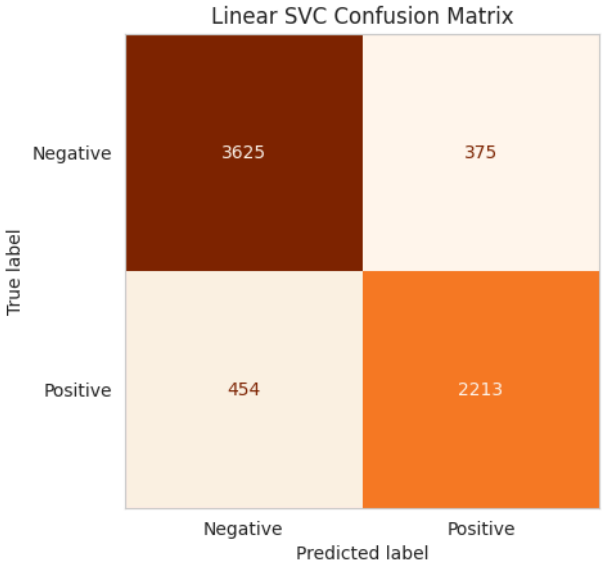


A2. TF-IDF — Imbalanced Classes

- Logistic Regression — Confusion Matrix

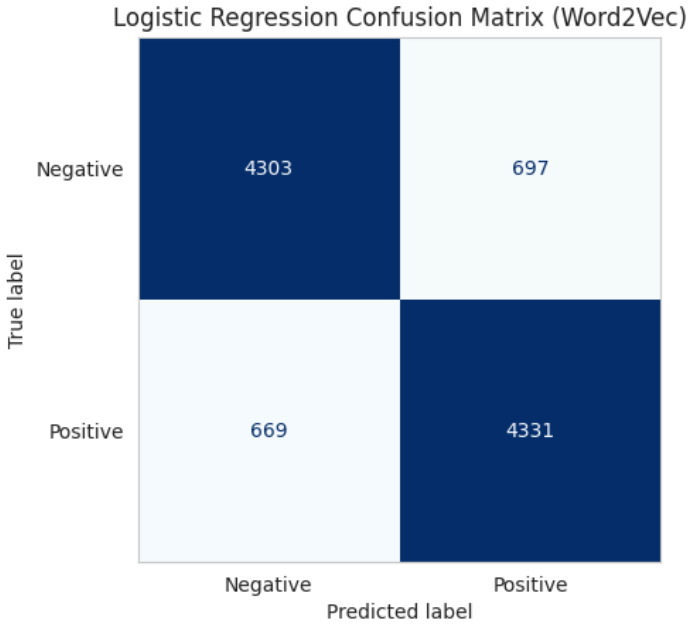


- Linear SVC — Confusion Matrix

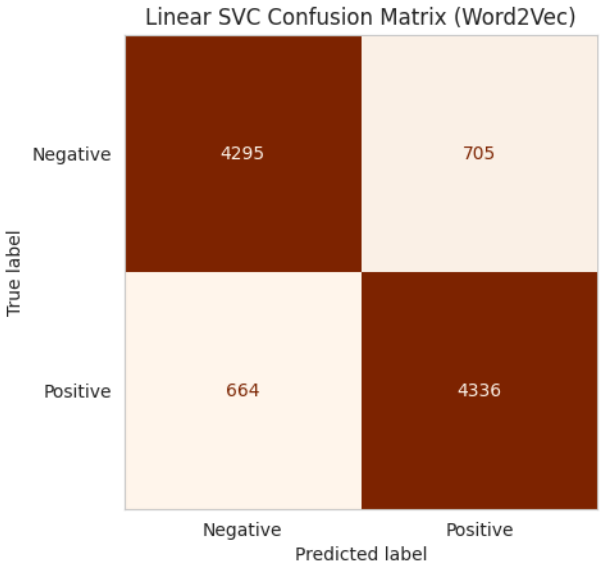


A3. Word2Vec — Balanced Classes

- Logistic Regression — Confusion Matrix

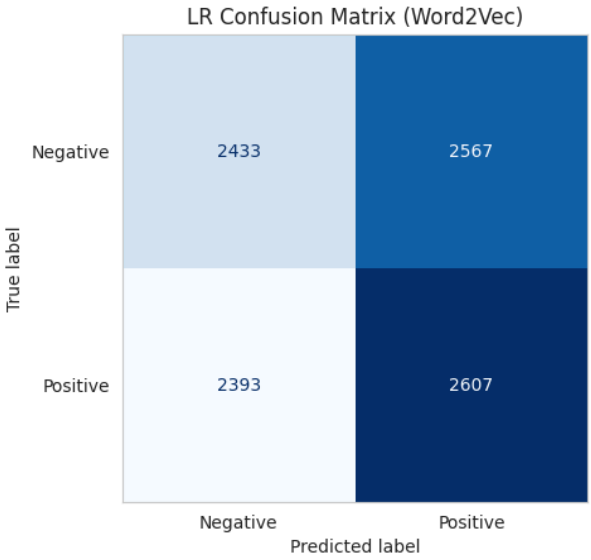


- Linear SVC — Confusion Matrix



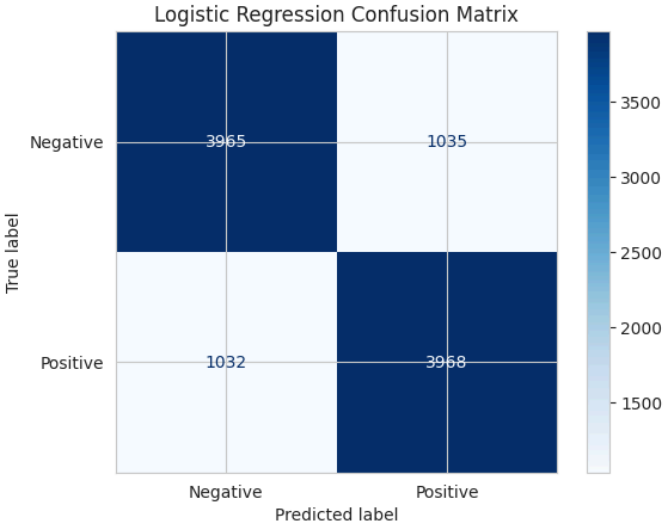
A4. Word2Vec — Imbalanced Classes

- Logistic Regression — Confusion Matrix

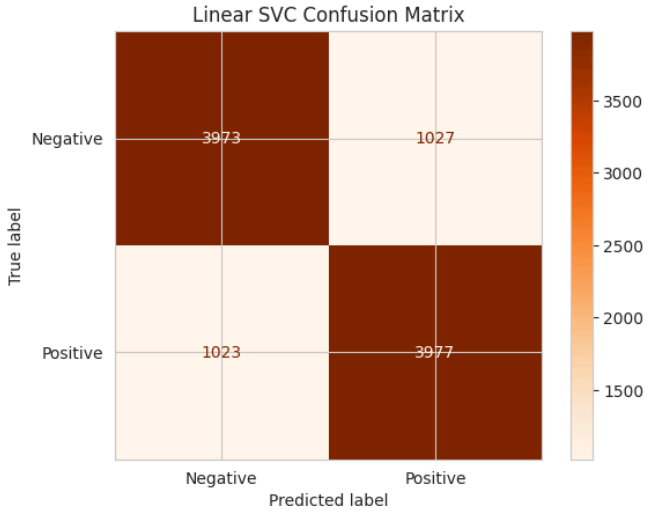


A5. GloVe — Balanced Classes

- Logistic Regression — Confusion Matrix

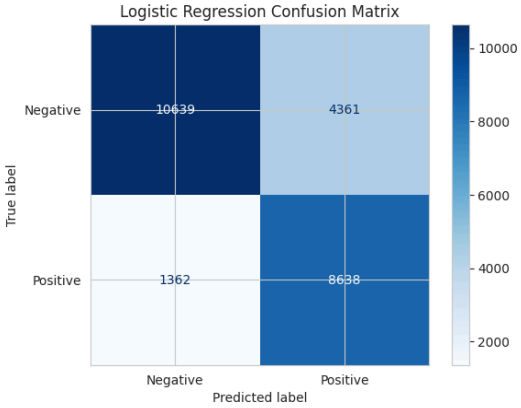


- Linear SVC — Confusion Matrix

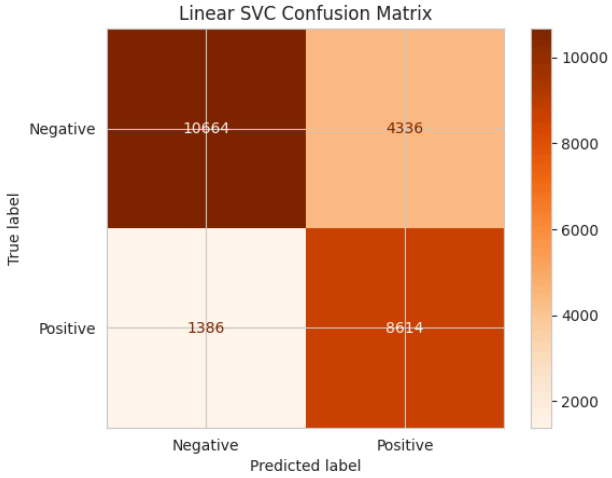


A6. GloVe — Imbalanced Classes

- Logistic Regression — Confusion Matrix



- Linear SVC — Confusion Matrix



Appendix B — Word Clouds Visualizations

B1. Word Cloud: Most Common Words in Positive Reviews

B2. Word Cloud: Most Common Words in Negative Reviews