Review Sentiment Analysis

using Text Classification

Christina Tzortzaki

Introduction

Sentiment analysis determines if a text os positive or negative.

Useful for business to analyze customer feedback.

Goal: Build and evaluate machine learning models for sentiment classification.

Data Collection

IMDB Movie Reviews (Training Dataset)

Size: 50,000 reviews (Balanced: 25K Positive | 25K Negative)
Columns: review (text), sentiment (1 = Positive, 0 = Negative)

Amazon Reviews (Testing Dataset)

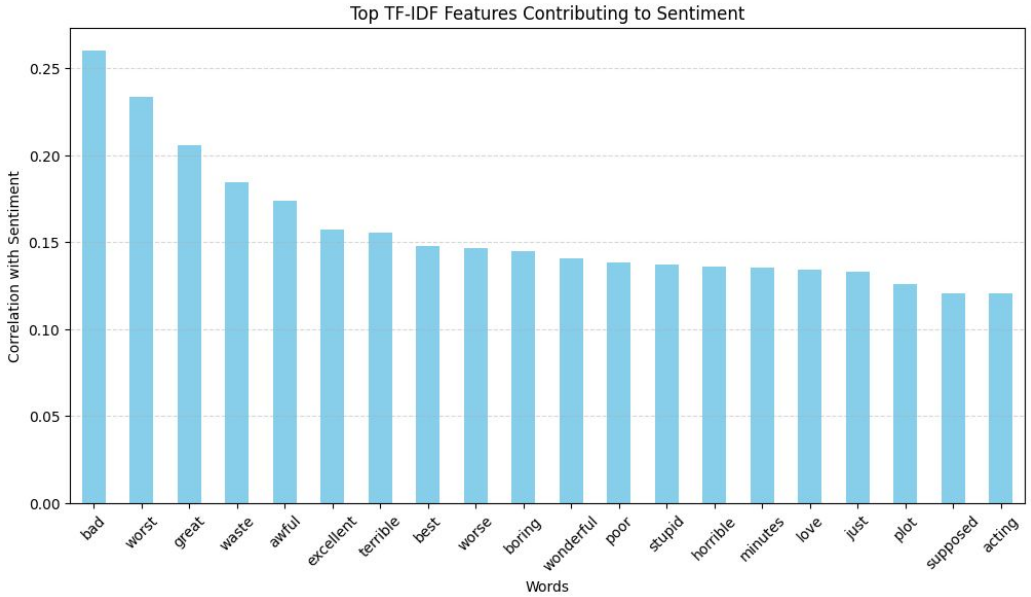
Size: ~200,000 reviews
Columns: review_text (text), classsss_index (1 = Negative, 2 = Positive), title (text)

Data Processing

- Removed HTML tags, punctuation and special characters.
- Converted text to lowercase.
- Applied ***TF-TIDF*** vectorizations (5000 max features) to transform text into numerical features.

Feature Exploration

Identifies top ***TF-IDF*** features contributing to sentiment.
Analyzed word importance using correlation.



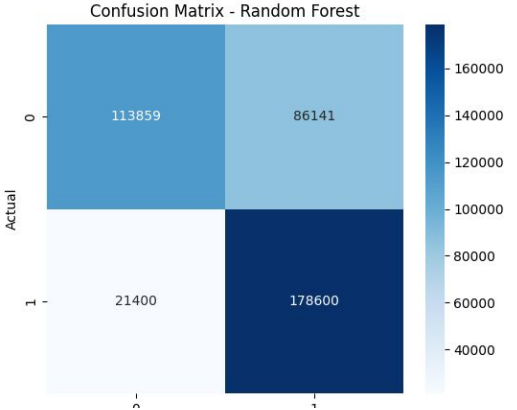
Machine Learning Models Used

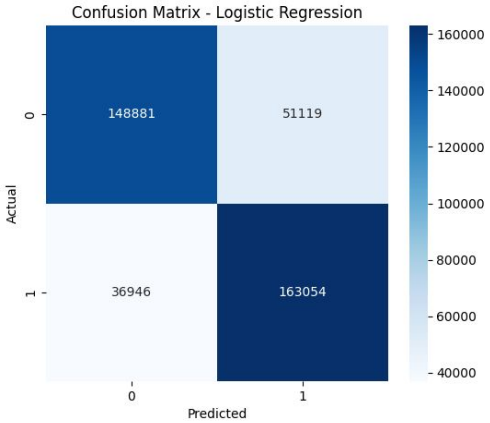
- Logistic Regression
- Naive Bayes
- Support Vector Machine (SVM)
- Random Forest

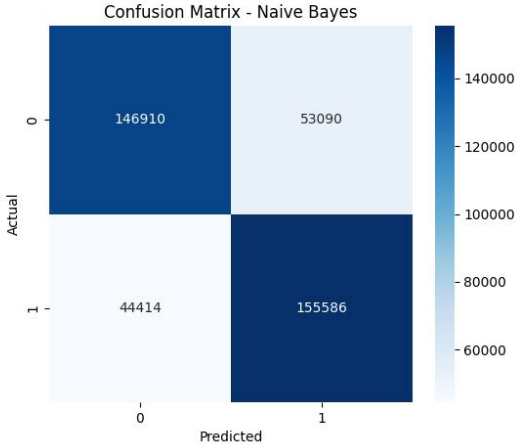
Each model was trained and evaluated for sentiment classification.

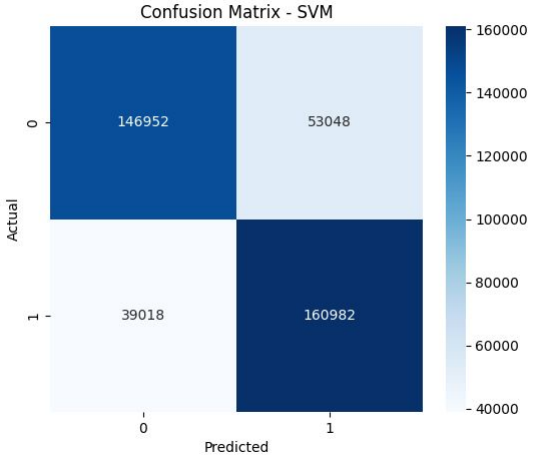
Model Training & Evaluation

Trained Data: IMDB Dataset.
Testing Data: Amazon Reviews





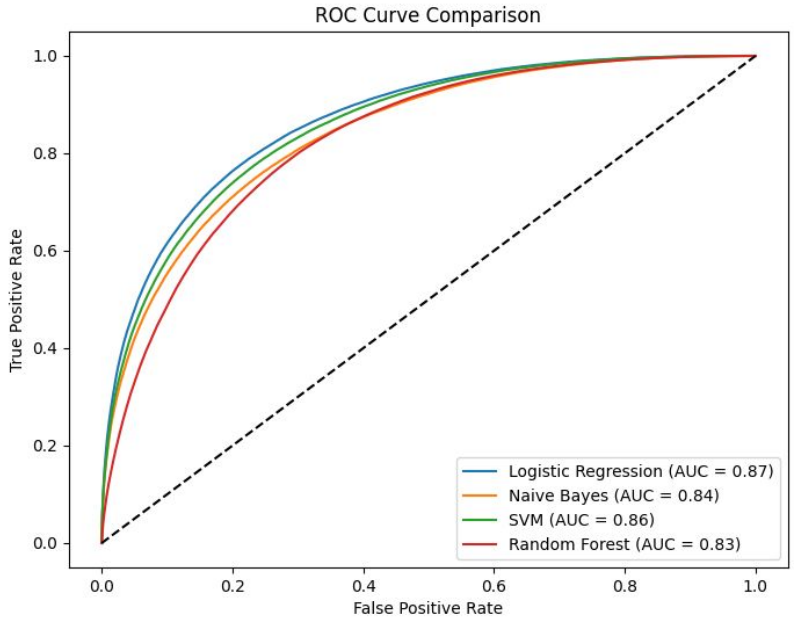




Receiver Operating Characteristic (ROC) Curve

Shows the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity) across different threshold values.

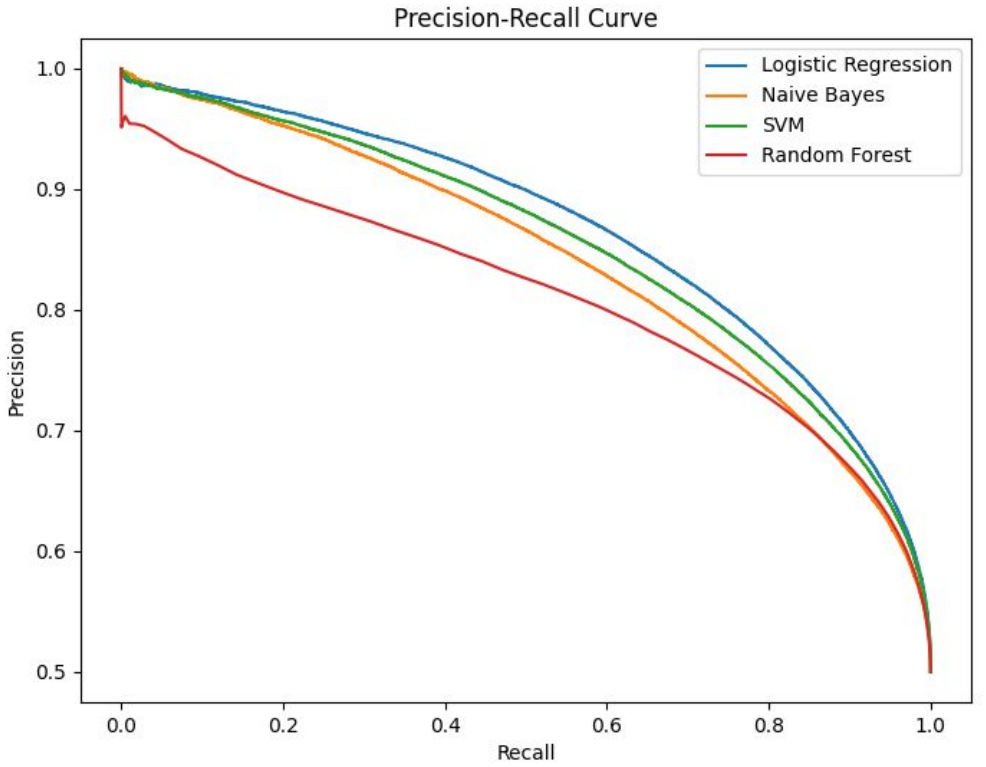
The dashed diagonal line (y = x) represents random guessing (AUC = 0.5).



Precision-Recall Curve

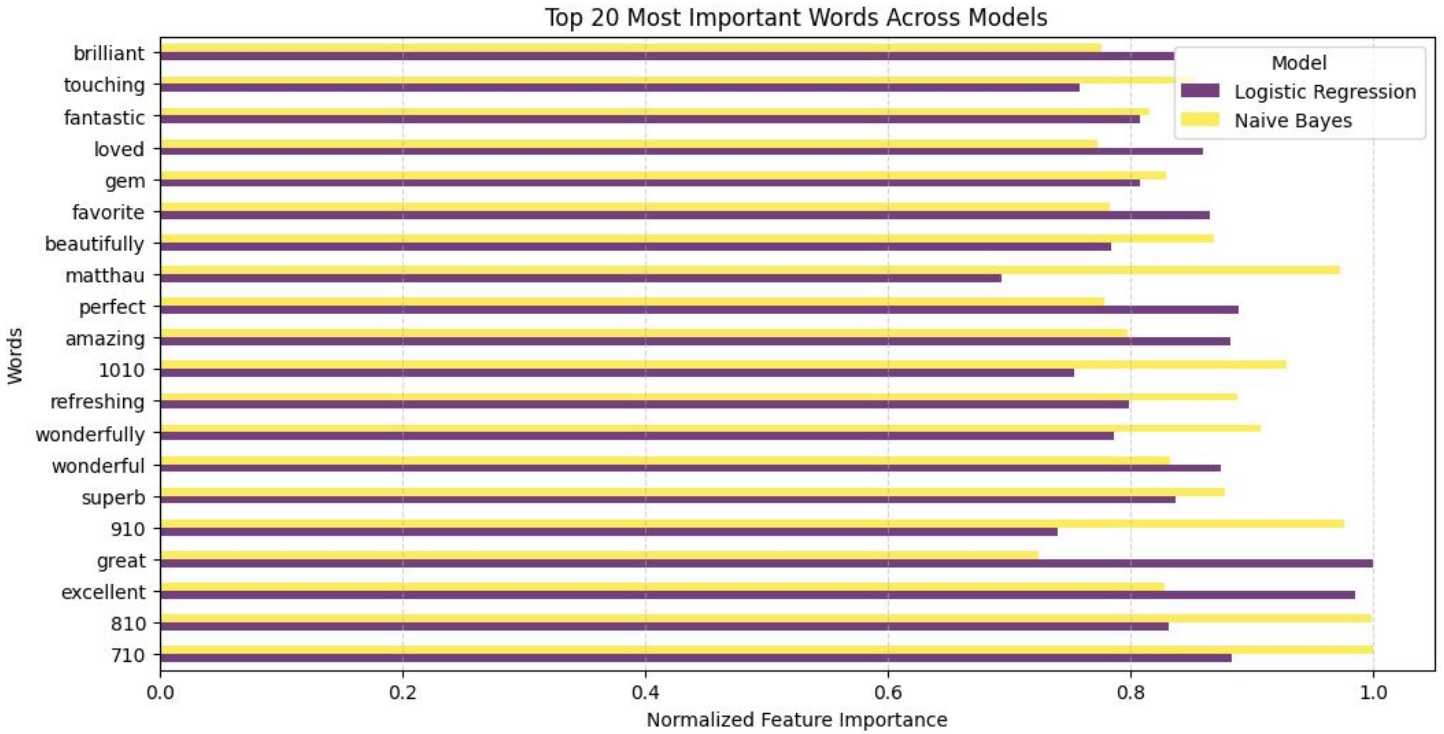
Compared precision-recall trade-offs across models.

Important for understanding model performance on imbalance data.



Evaluation metrics of Models

Models	F1-score (Class 0)	F2-score Class 1)	Accuracy	Recall (Class 0)	Recall (Class 1)
Logistic Regression	0.77	0.79	0.78	0.74	0.82
Naive Bayes	0.75	0.76	0.76	0.73	0.78
SVM	0.76	0.78	0.77	0.73	0.80
Random Forest	0.68	0.77	0.73	0.57	0.89


Top 20 Most Important Words Across Models
brilliant
touching
fantastic
loved
gem
favorite
beautifully
matthau
perfect
amazing
1010
refreshing
wonderfully
wonderful
superb
910
great
excellent
810
710
Words
0.0
0.2
0.4
0.6
0.8
1.0
Normalized Feature Importance
Model
Logistic Regression
Naive Bayes

Conclusion

- Logistic Regression is the best model overall, offering a balance between precision, recall, and F1-score.
- Random Forest is best for detecting Class 1 (high recall) but performs poorly for Class 0.
- Naïve Bayes and SVM perform similarly to Logistic Regression but with slightly lower scores.
- Sentiment analysis using machine learning is effective.

Thank you!

Any Questions?