IMDB Sentiment Analysis

Executive Summary

This project implemented a comprehensive sentiment analysis system for movie reviews using the IMDB dataset. Three machine learning models were trained and evaluated: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). The system achieved high performance across all models, with the best performing model reaching over 85% accuracy in classifying movie reviews as positive or negative sentiment.

Project Approach

Data Processing Pipeline

The project began with loading the IMDB dataset containing movie reviews and their corresponding sentiment labels. A robust text preprocessing pipeline was implemented to clean and prepare the textual data:

- **Text Normalization**: Converting all text to lowercase for consistency
- **HTML and URL Removal**: Eliminating web markup and hyperlinks that don't contribute to sentiment
- **Special Character Filtering**: Retaining only alphabetic characters and spaces
- **Stopword Elimination**: Removing common English words (the, and, or, etc.) that carry minimal sentiment information
- **Short Word Filtering**: Excluding words with fewer than 3 characters to reduce noise

This preprocessing step was crucial for improving model performance by focusing on meaningful content words that carry sentiment information.

Feature Engineering

The cleaned text was converted into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with the following configuration:

- Maximum of 10,000 features to balance performance and computational efficiency
- Both unigrams and bigrams (1-2 word sequences) to capture context
- Minimum document frequency of 2 to filter rare terms
- Maximum document frequency of 95% to exclude overly common terms
- Sublinear TF scaling to reduce the impact of term frequency

Model Selection and Training

Three complementary machine learning algorithms were selected to provide diverse approaches to the classification problem:

1. **Logistic Regression**: A linear model that provides interpretable coefficients and probability estimates
2. **Naive Bayes (Multinomial)**: A probabilistic classifier particularly effective for text classification
3. **Support Vector Machine (Linear)**: A robust classifier that finds optimal decision boundaries

All models were trained on 80% of the data with stratified sampling to maintain balanced class distribution across training and test sets.

Technical Challenges and Solutions

Data Quality Issues

Challenge: The raw dataset contained HTML tags, URLs, and various formatting artifacts that could mislead the models.

Solution: Implemented a comprehensive text cleaning function using regular expressions to remove HTML markup, URLs, and special characters while preserving meaningful content.

Feature Dimensionality

Challenge: Text data naturally creates high-dimensional sparse feature spaces that can lead to overfitting and computational inefficiency.

Solution: Applied TF-IDF vectorization with carefully tuned parameters to limit features to 10,000 most informative terms while using n-grams to capture contextual information.

Model Interpretability

Challenge: Understanding which features drive sentiment predictions for model validation and debugging.

Solution: Extracted and visualized the most influential features from the Logistic Regression model, showing both positive and negative sentiment indicators.

Empty Reviews Handling

Challenge: Some reviews became empty after aggressive text cleaning, which could cause errors during training.

Solution: Implemented validation to identify and remove reviews that contained no meaningful text after preprocessing.

Results and Performance Analysis

Model Performance Metrics

All three models demonstrated strong performance on the test dataset:

- **Accuracy**: All models achieved accuracy scores above 85%
- **F1-Score**: Balanced precision and recall metrics confirmed robust performance
- **Classification Reports**: Detailed per-class metrics showed balanced performance across positive and negative sentiments

Model Comparison

The evaluation revealed that while all models performed well, each had distinct characteristics:

- Logistic Regression provided the best balance of performance and interpretability
- Naive Bayes showed excellent performance with minimal computational overhead
- SVM demonstrated robust classification boundaries with strong generalization

Feature Analysis

Analysis of the Logistic Regression coefficients revealed meaningful sentiment indicators:

- **Positive indicators**: Words like "excellent," "amazing," "brilliant," "outstanding"
- **Negative indicators**: Words like "terrible," "awful," "boring," "waste"
- **Contextual phrases**: Bigrams captured more nuanced expressions like "highly recommend" vs "completely disappointed"

Practical Implementation

Interactive Analysis System

The project includes a user-friendly interface allowing real-time sentiment analysis of custom movie reviews. Key features include:

- Multi-model consensus scoring
- Confidence levels for predictions
- Detailed probability breakdowns
- Batch processing capabilities

Prediction Functionality

The system provides comprehensive prediction capabilities:

- Individual review analysis with cleaning pipeline visualization
- Confidence scores and probability distributions
- Model consensus reporting
- Export functionality for batch analysis results

Conclusions and Impact

The IMDB sentiment analysis system successfully demonstrates the effectiveness of traditional machine learning approaches for text classification. The high accuracy achieved across all three models validates the preprocessing and feature engineering choices made during development.

Key Achievements

1. **High Accuracy**: Achieved over 85% accuracy across all models
2. **Robust Preprocessing**: Developed a comprehensive text cleaning pipeline
3. **Model Diversity**: Implemented multiple algorithms for comparison and consensus
4. **Practical Usability**: Created an interactive system for real-world application
5. **Interpretability**: Provided insights into feature importance and model decision-making

Future Enhancements

While the current system performs excellently, several areas offer opportunities for improvement:

- Integration of modern transformer-based models (BERT, RoBERTa) for potentially higher accuracy
- Advanced preprocessing techniques like lemmatization and named entity recognition
- Ensemble methods combining all three models for improved performance
- Extension to multi-class sentiment analysis (very negative, negative, neutral, positive, very positive)
- Real-time streaming analysis capabilities for live review processing

The project demonstrates that with careful data preprocessing, appropriate feature engineering, and thoughtful model selection, traditional machine learning approaches can achieve highly effective sentiment analysis for practical applications.