IMDb Sentiment Analysis Report

1. Dataset Source and Summary

The dataset used in this project is the ACL IMDb movie reviews dataset, containing 50,000 labeled reviews divided equally into training and test sets. Each review is labeled as either positive or negative, making this a binary classification task. The dataset is balanced and widely used for sentiment analysis benchmarking.

2. Process Steps

- Preprocessing: HTML tag removal, tokenization, lemmatization.
- TF-IDF Feature Extraction: Using both unigrams and bigrams (1,2).
- Class-Specific Analysis: Separate TF-IDF vectors were calculated for positive and negative reviews.
- Discriminative Phrases: Calculated difference in TF-IDF values to identify top n-grams unique to each class.
- Visualizations: WordClouds of top positive and negative n-grams.
- Modeling: Logistic Regression and Naive Bayes classifiers trained on the TF-IDF vectors.

3. Model Results

- Logistic Regression Accuracy: 88%
- Naive Bayes Accuracy: 86%

Both models performed well, with logistic regression slightly outperforming Naive Bayes.

4. Insights & Limitations

- Some common terms still appear across both classes (e.g., 'movie', 'film'), but filtering using n-gram differences greatly improved class separation.
- Context-aware n-grams such as 'waste time' and 'highly recommend' were crucial for model performance.
- Simple models like logistic regression can perform surprisingly well with carefully selected features.

5. Key Insights Learned

- Context is critical for sentiment: unigrams alone are not enough.
- Discriminative n-gram extraction (TF-IDF difference) is powerful for interpretable models.
- Visualization helps validate feature selection and model logic effectively.