IMDb Movie Review Sentiment - Final Report

1. Problem Statement

The aim of this project is to build a machine learning model that can classify IMDb movie reviews as **positive** or **negative**. The dataset consists of 50,000 labeled movie reviews. The objective is to apply text preprocessing, feature engineering, and machine learning techniques to achieve a high-performing sentiment classification model.

2. Data Exploration

- Dataset: 50,000 reviews with equal distribution between positive and negative labels.
- Visuals:
    - Sentiment distribution (balanced).
    - Review length distribution plotted to understand text characteristics.

3. Data Preprocessing

Steps:

- Removed HTML tags and non-alphabetic characters.
- Converted text to lowercase.
- Tokenized the reviews.
- Removed stopwords.
- Applied lemmatization.

Output Column: clean_review containing preprocessed reviews.

4. Feature Engineering

- **Word Count**, **Character Count**, **Average Word Length** were computed.
- TF-IDF vectorization applied to transform text into numerical features.
- max_features=5000 used in TfidfVectorizer.

5. Model Development & Evaluation

A. Random Forest Classifier

- **Hyperparameters Tuned**: n_estimators, max_depth, min_samples_split
- **Best Parameters**: n_estimators=75, max_depth=15, min_samples_split=2
- **Accuracy**:
  - Train: 88.6%
  - Test: **82.9%**
- **F1-score**: 83%

B. Logistic Regression

- **Hyperparameters Tuned**: C, solver
- **Best Parameters**: C=2.91, solver='liblinear'
- **Accuracy**:
  - Train: 92.3%
  - Test: **88.7%**
- **F1-score**: 89%
- **Best Performing Model**

C. Support Vector Machine (LinearSVC)

- No hyperparameter tuning.
- **Accuracy**:
  - Train: 92.9%
  - Test: **88.2%**
- **F1-score**: 88%

D. LSTM (Deep Learning Model)

- Used Keras with:
  - Embedding → LSTM → Dropout → Dense
  - maxlen=200, vocab_size=10000
- **Accuracy**:
  - Test: **88.4%**
- **F1-score**: ~88%
- Implemented based on self-learning, not covered in lectures.

6. Visualization

- **Word Clouds** generated for both positive and negative reviews to visualize frequently used terms.
  - Positive reviews: Clean, hopeful, emotional terms.
  - Negative reviews: Critical, disappointed tone.

7. Future Predictions

- **Custom Function** built to allow real-time sentiment prediction using the trained Logistic Regression model and TF-IDF transformation.
- Predicts sentiments for new input reviews interactively.

8. Conclusion

Model	Test Accuracy	Comments
Logistic Regression	**88.7%**	Best performer with optimal F1-score
LSTM	88.4%	Strong performance using deep learning
SVM (LinearSVC)	88.2%	Competitive accuracy
Random Forest	82.9%	Lowest among the tested models

Final Choice: Logistic Regression was selected for future predictions due to its balance of performance and simplicity.

Video Explanation Link:

https://drive.google.com/file/d/1nw_0co-exFVrcuq8I-XDfWX-QMBBm3tF/view?usp=sharing