Movie Sentiment Analysis Project Report

1. Introduction

The Movie Sentiment Analysis project aims to classify user reviews as positive or negative using Natural Language Processing (NLP) and Machine Learning techniques. This system can be used to gauge audience reactions and opinions on movies, aiding content creators and platforms.

2. Approach Used

Data Collection and Preprocessing

- **Dataset**: IMDB movie review dataset was used, containing labeled movie reviews.
- **Preprocessing steps**:
    - Lowercasing
    - Removal of punctuation and special characters
    - Tokenization and stop word removal
    - Lemmatization/Stemming
    - Vectorization using TF-IDF and Word Embeddings (for LSTM)

Model Training

Three types of models were trained and evaluated:

- **Naïve Bayes**: Fast and interpretable baseline model, using TF-IDF vectors.
- **Random Forest**: Ensemble method to capture nonlinear relationships in the data.
- **LSTM (Long Short-Term Memory)**: Deep learning approach that captures sequence-based dependencies using word embeddings.

3. Challenges Faced

- **Data Imbalance**: Some versions of datasets had slightly skewed sentiment classes.
- **Overfitting**: Especially with LSTM when training on small datasets or too many epochs.
- **Text Noise**: Handling sarcasm, misspellings, and informal language in reviews.
- **Model Deployment**: Hosting deep learning models with limited backend resources.

4. Model Performance & Improvements

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naïve Bayes	**83%**	**82%**	**84%**	**83%**
Random Forest	**85%**	**86%**	**84%**	**85%**
LSTM	**88%**	**87%**	**89%**	**88%**

Improvements Made

- Implemented dropout and regularization in LSTM to avoid overfitting.
- Used GloVe embeddings for better semantic understanding.
- Hyperparameter tuning (learning rate, max depth, n_estimators) for Random Forest.