# Movie Review Sentiment Analysis

**1. Project Overview**

This project introduces interns to fundamental concepts in natural language processing (NLP) and machine learning. By the end of the task, participants will have built and evaluated a sentiment classifier using real-world text data from the IMDb Movie Reviews Dataset.

**2. Project Objectives**

* Apply standard NLP techniques to clean and preprocess movie review data.
* Train a machine learning classifier capable of detecting sentiment in text.
* Evaluate the model performance using accuracy and F1-score.
* (Optional) Deploy a simple user interface to demonstrate the sentiment prediction system.

**3. Methodology**

**1. Text Preprocessing**

The following steps will be applied to clean and standardize the review text:

* **Lowercasing**: Convert all text to lowercase for uniformity.
* **Removing Stopwords**: Filter out common English words (e.g., “the”, “is”) that do not contribute to sentiment.
* **Tokenization**: Split reviews into individual words or tokens.
* **(Optional Enhancements)**: Lemmatization or stemming to reduce words to their root forms.

**2. Feature Extraction**

Convert text into numerical format for model training:

* **Bag of Words (BoW)** or **TF-IDF Vectorization** for word representation.
* The resulting vectors are used as input features for the machine learning model.

**3. Model Training**

Choose and implement one or more of the following classification algorithms:

* **Logistic Regression**: A strong baseline for binary classification.
* **Naïve Bayes**: Effective for text classification with independent feature assumption.
* **Support Vector Machine (SVM)**: Useful for high-dimensional text data.

**4. Model Evaluation**

* Use metrics to evaluate the classifier’s performance:
  + **Accuracy**: Overall correctness of the model.
  + **F1-Score**: Harmonic mean of precision and recall, especially important if there's class imbalance.
* Perform k-fold cross-validation to test generalizability.

**4. Optional: Interface Development**

To enhance usability:

* **Build a simple web interface** using tools like Streamlit or Flask.
* Allow users to input a custom movie review.
* Display the predicted sentiment as **Positive** or **Negative** in real time.

**5. Tools & Technologies**

* **Programming Language**: Python
* **Libraries**:
  + Text Preprocessing: NLTK, spaCy
  + Feature Extraction: scikit-learn (TfidfVectorizer, CountVectorizer)
  + Modeling: scikit-learn (LogisticRegression, MultinomialNB, SVM)
  + Interface (Optional): Streamlit or Flask