**Major Project 2:**

**NLP Application**

**Task Selection: Sentiment Analysis**

**Objective:** To classify movie reviews as positive or negative.

**Data Preprocessing**

1. **Data Acquisition:** Load the IMDB movie review dataset.
2. **Text Cleaning:** Remove stop words, punctuation, and convert text to lowercase.
3. **Tokenization:** Split text into individual words or tokens.
4. **Text Representation:** Use TF-IDF to convert text into numerical vectors.

**Model Selection**

* **Random Forest Classifier:** A popular ensemble learning algorithm for classification tasks.

**Model Training**

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, f1\_score

# Load dataset (replace with your data)

X = ... # List of movie reviews

y = ... # Corresponding sentiment labels (0 or 1)

# Preprocessing

vectorizer = TfidfVectorizer()

X\_vectorized = vectorizer.fit\_transform(X)

# Split data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_vectorized, y, test\_size=0.2, random\_state=42)

# Create and train the model

model = RandomForestClassifier()

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model.predict(X\_test)

# Evaluate performance

accuracy = accuracy\_score(y\_test, y\_pred)

f1 = f1\_score(y\_test, y\_pred)

print("Accuracy:", accuracy)

print("F1-score:", f1)

**Results Presentation**

* Discuss the model's performance based on accuracy and F1-score.
* Analyze the model's strengths and weaknesses.
* Explore potential improvements, such as using different models or preprocessing techniques.

**Additional Considerations:**

* **Deep Learning Models:** For more complex tasks, consider using deep learning models like LSTM or Transformer.
* **Evaluation Metrics:** Choose appropriate evaluation metrics based on your task (e.g., BLEU score for machine translation, ROUGE score for summarization).
* **Data Augmentation:** If your dataset is small, consider techniques like data augmentation to increase its size.