**Part A: IMDB Sentiment Analysis**

**Sentiment analysis is a key application in Natural Language Processing (NLP) that helps determine the sentiment of text data. This project focuses on classifying IMDB movie reviews as positive or negative using various Machine Learning (ML) and Deep Learning models.**

**1. Data Exploration**

**Dataset Overview**

* **The dataset contains 50,000 IMDB movie reviews labelled as "positive" or "negative."**
* **No missing values were found in the dataset.**
* **The distribution of positive and negative reviews is balanced.**

**Data Visualization**

* **A histogram was generated to analyse the distribution of review lengths.**

**2. Data Preprocessing**

* **HTML tags and non-alphabetic characters were removed.**
* **Stop words were filtered out, and words were lemmatized and stemmed.**
* **Reviews were vectorized using Count Vectorizer, TF-IDF, and tokenized for deep learning models.**

**3. Model Development**

**Machine Learning Models Used:**

* **Logistic Regression**
* **Naive Bayes**
* **Support Vector Machine (SVM)**
* **Random Forest**

**Deep Learning Models Used:**

* **Long Short-Term Memory (LSTM)**
* **BERT (Bidirectional Encoder Representations from Transformers)**

**5. Model Evaluation**

**All models were evaluated using accuracy, F1-score, and ROC-AUC score.**

**Comparison Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model |  | Accuracy |  | F1-Score |  | ROC-AUC Score |
| Logistic Regression |  | 88% |  | 89% |  | 88% |
| Naive Bayes |  | 85% |  | 85% |  | 85% |
| SVM |  | 87% |  | 88% |  | 87% |
| Random Forest |  | 84% |  | 84% |  | 84% |
| LSTM  BERT |  | 87%  - |  | -  - |  | -  - |

**6. Best Model Selection**

**Based on evaluation metrics, Logistics Regression performed the best, achieving the highest accuracy, F1-score, and ROC-AUC score.**

**7. Insights and Future Recommendations**

**Insights:**

* **ML model (Logistics Regression) outperformed traditional ML models.**
* **TF-IDF representation worked well for ML models.**
* **Training time significantly increased for complex models like SVM and BERT.**

**Future Recommendations:**

* **Experiment with hyperparameter tuning for ML models.**
* **Implement data augmentation techniques for better generalization.**
* **Deploy the best-performing model as a web service for real-world applications.**

**8. Conclusion**

**This project successfully classified IMDB reviews using multiple models. Logistic Regression emerged as the ML model.**

**Part B: News Article Classification Report**

**1. Introduction**

**This report presents the process of classifying news articles into different categories using machine learning techniques. The project includes data collection, preprocessing, feature extraction, model training, and evaluation.**

**2. Data Collection**

**The dataset consists of labelled news articles from various categories such as sports, politics, and technology. It includes features like headlines, short descriptions, and keywords**.

**3. Data Preprocessing**

**To prepare the dataset for model training, the following preprocessing steps were applied:**

* **Text Cleaning: Removing special characters and converting text to lowercase.**
* **Tokenization:** **Splitting text into words.**
* **Stopword Removal:** **Eliminating common words that do not contribute to meaning.**
* **Lemmatization:** **Reducing words to their base forms.**

**4. Feature Extraction**

**To convert text data into numerical representations, Word2Vec word embeddings were used. The embeddings capture the semantic meaning of words and provide meaningful feature vectors**.

**5. Model Development**

Three machine learning models were trained for classification:

* **Naive Bayes Classifier**
* **Logistic Regression**
* **Support Vector Machine (SVM)**

**The dataset was split into training and testing sets using an 80-20 split. The models were trained and evaluated on the processed data.**

**6. Model Evaluation**

The models were evaluated using the following performance metrics:

* **Accuracy**
* **Precision**
* **Recall**
* **F1 Score**

**Performance Summary:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Naive Bayes | **36%** | **36%** | **36%** | **35%** |
| Logistic Regression | **54%** | **54%** | **54%** | **54%** |
| SVM | **55%** | **56%** | **55%** | **55%** |

**7. Visualization**

**Visualizations were created to understand the data and model performance:**

* **Category Distribution: A bar chart showing the frequency of each category.**
* **Confusion Matrices: To visualize model predictions versus actual values.**

**8. Conclusion**

**The trained models successfully classify new, unseen news articles into predefined categories. Among the models tested, the best-performing model can be selected based on evaluation metrics**.

**9. Future Work**

**To further improve classification accuracy, the following enhancements can be explored:**

* **Using advanced embeddings like BERT for better feature extraction.**
* **Hyperparameter tuning for improved model performance.**
* **Implementing deep learning models such as LSTMs or transformers.**

**This concludes the report on news article classification. The findings demonstrate the effectiveness of machine learning in text classification tasks.**