Team Infinity – Serve Smart Hackathon BHU

Report on Model Development and Results

Introduction

This report details the development and results of a machine learning model aimed at [objective, e.g., classifying news articles as real or fake]. The work encompasses data preprocessing, exploratory data analysis, model training, evaluation, and testing. Two primary notebooks, **main.ipynb** and **model.ipynb**, document the implementation.

Dataset Description

The dataset used is a tab-separated file, "test.tsv," which contains labeled samples. Key attributes include:

- **Text**: The main content.
- **Label**: The classification label (e.g., 0 for fake, 1 for real).

Initial exploration revealed the dataset to be balanced, as illustrated by label distribution plots.

Methodology

1. Data Preprocessing

- Libraries used: Pandas, NumPy, NLTK, spaCy, Matplotlib, Seaborn.
- Preprocessing steps included:
    - Loading the dataset.
    - Visualizing label distributions to assess class balance.
    - Text normalization (lowercasing, removing special characters, tokenization, stopwords removal, and lemmatization).

2. Exploratory Data Analysis (EDA)

- Visualizations were created to analyze text length, word distributions, and correlation between text features and labels.
- Example: A Seaborn count plot illustrated the balance between labels.

3. Model Training

- Models developed: Logistic Regression, Random Forest
- Feature engineering: TF-IDF vectorization applied to text data.
- Training and validation involved stratified sampling to preserve label distributions.

4. Evaluation

- Metrics: Accuracy, precision, recall, F1-score.
- Cross-validation performed to ensure robustness.

Results

The model achieved the following performance metrics on the test set:

- **Accuracy**: 96
- **F1-score**: 96
- **Precision:** 95
- **Recall:** 96
- **AUC ROC:** 99

Visualizations and confusion matrices highlighted the model's strengths and areas for improvement.

Conclusion

This project demonstrates a structured approach to solving a text classification problem and provides a strong foundation for further improvements.