Explainable AI – Lab Assignment Report

Course: 23CA201SE402 – Explainable AI (P)
Batch/Section: 23CSBTB37
Roll No.: 2303A52076

Objective

To apply Explainable AI (XAI) techniques using LIME (Local Interpretable Model-Agnostic Explanations) on text classification tasks:

1. IMDB Movie Review Sentiment Analysis
2. Fake vs Real News Detection

Models used: TF-IDF Vectorization with Logistic Regression.

Datasets

1. IMDB Dataset.csv – 50,000 reviews labeled positive/negative.
2. Fake.csv and True.csv – merged dataset of fake vs real news articles.

Problem 1: Sentiment Analysis (IMDB)

- **Preprocessing**: Converted sentiment to binary (positive=1, negative=0).
- **Model**: Logistic Regression with TF-IDF features.
- **Accuracy**: 0.8889 (~88.9%)
- **Example Review (truncated)**:
  "I've watched this movie on a fairly regular basis for most of my life..."
- **Predicted Class**: Positive (Prob: 0.90 vs 0.10 negative)
- **LIME Explanation (Top Positive Words)**: *Tommy, best, funny, friends, great*

Problem 2: Fake News Detection

- **Preprocessing**: Merged Fake.csv and True.csv, added labels (Fake=1, Real=0).
- **Model**: Logistic Regression with TF-IDF features.
- **Accuracy**: 0.9861 (~98.6%)
- **Example Article (truncated)**:
  "WASHINGTON (Reuters) – New Jersey Governor Chris Christie..."
- **Predicted Class**: REAL (Prob: 0.96 vs 0.04 fake)
- **LIME Explanation (Top Influential Words)**: *Reuters, Christie, Washington, said, president*

Observations

- IMDB sentiment achieved ~89% accuracy, proving Logistic Regression with TF-IDF is effective for short reviews.
- Fake News detection reached ~99% accuracy, showing strong separability between fake and real articles.
- LIME provided clear word-level explanations, making model predictions more transparent.
- Influential words matched human intuition (e.g., *funny* in positive reviews, *Reuters* in credible news).

Conclusion

Both experiments demonstrate the usefulness of LIME in making text classification models interpretable. While Logistic Regression already has interpretable coefficients, LIME gives instance-level explanations that highlight why a specific prediction was made.