

Name – Shruthi Addagudi

2303A52133 Batch – 37

Report for Explainable AI – Assignment 3

Problem 1: Sentiment Analysis with LIME

Problem Statement

The task was to perform sentiment classification on a given text dataset and to use **LIME (Local Interpretable Model-agnostic Explanations)** to understand which words influence the model's predictions the most.

Steps Followed

1. Data Loading

- The dataset was loaded, containing text samples labeled as positive or negative.

2. Preprocessing

- Text was cleaned (removing punctuation, lowercasing, tokenization).
- Data was split into training and validation sets.

3. Feature Extraction

- Applied **TF-IDF Vectorization** to convert text into numerical features suitable for machine learning models.

4. Model Training

- Trained a **Logistic Regression classifier** to predict sentiment.
- Model achieved good accuracy on validation data.

5. Explainability with LIME

- Used LimeTextExplainer to interpret predictions.
- LIME highlighted the most important words contributing to a prediction.
- Example: For a positive review, words like *"great"*, *"amazing"* were shown as contributing positively, while for a negative review, words like *"bad"*, *"boring"* contributed negatively.

Observations

- Logistic Regression with TF-IDF provided interpretable and strong results.
- LIME explanations matched human intuition, showing that the model relies on sentiment-indicative words.

Conclusion

- The task successfully demonstrated **Explainable AI (XAI)** using LIME.

- Sentiment classification not only predicts but also provides transparency in decision-making.
-

Problem 2: Fake News Detection with LIME

Problem Statement

The task was to build a **Fake News Detection** model using the **FakeNews dataset**, where the goal is to classify news articles as *fake* or *legit*. The model should also use LIME to highlight suspicious words influencing predictions.

Steps Followed

1. **Data Loading**
 - Loaded dataset from folders:
 - fakeNewsDataset/fake/ → Fake news articles
 - fakeNewsDataset/legit/ → Real/legitimate news articles
2. **Preprocessing**
 - Combined text files into a labeled dataset (Fake = 0, Legit = 1).
 - Cleaned and tokenized text.
3. **Feature Extraction**
 - Applied **TF-IDF Vectorizer** to convert articles into numerical features.
4. **Model Training**
 - Trained a **Logistic Regression classifier**.
 - The model learned to differentiate fake vs real news with good performance.
5. **Explainability with LIME**
 - Applied LimeTextExplainer to individual predictions.
 - LIME highlighted suspicious words often found in fake news (e.g., "*shocking*", "*breaking*", "*claims*").
 - Legit news articles were associated with words like "*report*", "*official*", "*statement*".

Observations

- The Logistic Regression model was effective in separating fake vs real articles.
- LIME highlighted contextually meaningful suspicious words, which aligns with how humans judge news credibility.
- This increases **trust** and **interpretability** of the fake news detection system.

Conclusion

- Successfully developed a **fake news classifier with interpretability**.

- LIME explanations provided useful insights into which words contribute to a news article being considered fake or real.
-

Final Insights

- Both problems show the **importance of Explainable AI**.
- LIME makes models transparent, helping humans understand and trust machine learning predictions.
- Such techniques are crucial in **sensitive domains** like sentiment analysis for reviews and fake news detection in media.