CS 519 Open Machine Learning Project

1 Fake News Detection Using NLP and Machine Learning

1.1 Motivation

Misinformation and fake news have severe social and political consequences. Developing a fake news detector can help improve information credibility.

1.2 Problem

Given a dataset of news articles, classify whether a news article is real or fake.

1.3 Solution Approach

- **Text Preprocessing:** Tokenization, stopword removal, TF-IDF, and Word2Vec embeddings.
- **Supervised Learning Models:** Train classifiers like Logistic Regression, LSTMs, BERT, and Random Forests.
- **Explainability:** Use SHAP values to interpret why a model classifies certain news as fake.

1.4 Datasets

- Fake News dataset on UCI ICS.
- LIAR dataset (Politifact-labeled fake news).
- Fake News dataset on Kaggle.

Examples of available datasets:

- `https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`
- `https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification`
- `https://archive.ics.uci.edu/ml/datasets/News+Aggregator`
- `https://www.cs.ucsb.edu/~william/data/liar_dataset.zip`

We can select a similar dataset from these options.

1.5 Analysis

- Evaluate using F1-score, Precision, Recall.
- Compare different ML models on accuracy and interpretability.
- Test model robustness against adversarial samples (e.g., small textual changes in fake news headlines).

1.6 Team Members

- **Tathagata Debnath**
  Email: tathadbn@nmsu.edu
  Student ID: 800678028
- **Tomas Urena**
  Email: tomasu@nmsu.edu
  Student ID: 800756051
- **Reagan Rubio**
  Email: reygunr@nmsu.edu
  Student ID: 800799845
- **Sayantan Maity**
  Email: smrik110@nmsu.edu
  Student ID: 800812216