Fake News Classification using NLP

Introduction

This mini-project focuses on building a system to classify news articles as either *Reliable* or *Unreliable*. Using Natural Language Processing (NLP) techniques, machine learning, and Python libraries, the project implements a comprehensive pipeline from data preprocessing to deployment via a user-friendly interface using Streamlit and Ngrok.

Objectives

- Preprocess textual data to remove noise and extract meaningful features.
- Train a Decision Tree Classifier for fake news classification.
- Deploy the system using Streamlit for real-time accessibility.

Dataset

The dataset used (`train.csv`) contains the following columns:

- **text**: The content of the news article.
- **label**: The classification target where:
  - 0 indicates *Reliable*.
  - 1 indicates *Unreliable*.

Preprocessing Steps:

- Missing values were replaced with empty strings.

- Unnecessary columns such as `id`, `title`, and `author` were dropped.
- Text was cleaned by removing special characters, converting to lowercase, tokenizing, and removing stopwords.
- Words were normalized to their root forms using the **Porter Stemmer**.

Methodology

1. Data Preprocessing

Preprocessing steps included:

- Removing special characters using regular expressions.
- Tokenizing text and removing stopwords.
- Applying stemming to reduce words to their base forms.

2. Feature Extraction

Text data was converted to numerical features using the **TF-IDF Vectorizer**, a method that captures the importance of words in the document relative to the corpus.

3. Model Training

- A **Decision Tree Classifier** was trained on the TF-IDF vectorized data.
- The dataset was split into 80% training data and 20% testing data.

4. Deployment

The trained model and vectorizer were saved as `.pkl` files, and a simple interface was developed using **Streamlit**. The app was made accessible online using **Ngrok**.

Results

The Decision Tree Classifier was evaluated on both training and test data:

- **Training Accuracy:** 99%
- **Test Accuracy:** 88%

Example Prediction:

- *Input:* "In these trying times, Jackie Mason is the n this week's exclusive clip for Breitbart News..."
- *Prediction:* Unreliable

The model's performance was visualized using a confusion matrix and a learning curve:

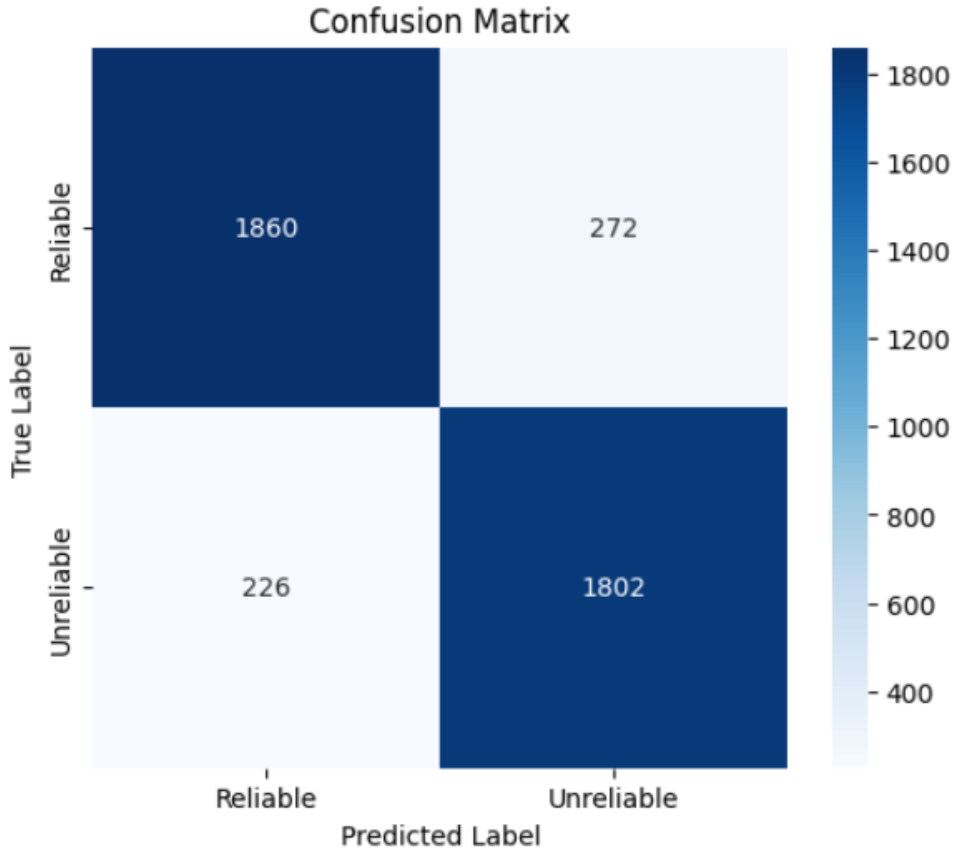

Figure 1: Confusion Matrix of Test Set Predictions

Tools and Libraries

- Python Libraries: `nltk`, `scikit-learn`, `pandas`, `matplotlib`, `seaborn`, `streamlit`, and `pyngrok`.

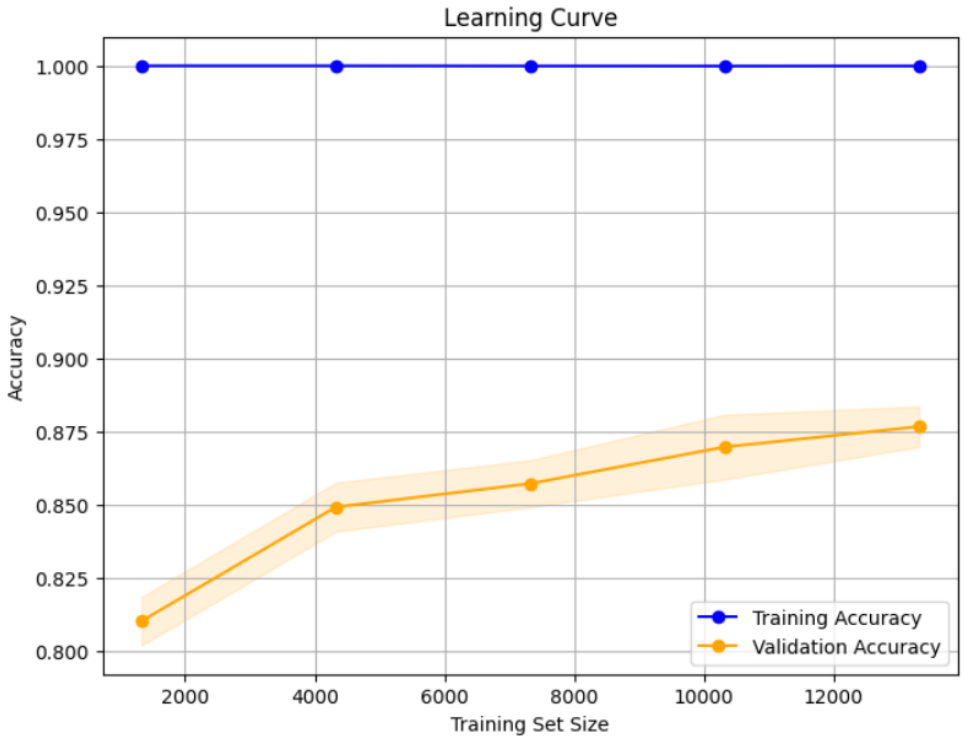

Figure 2: Learning Curve

- Machine Learning Algorithm: **Decision Tree Classifier**.
- Deployment Tools: **Streamlit** for UI and **Ngrok** for hosting.

Conclusion

This project showcases the application of NLP techniques and machine learning to solve the problem of fake news detection. The system effectively preprocesses text data, extracts features, trains a classifier, and provides an intuitive interface for real-time predictions.

Future Enhancements

- Incorporate advanced models like **BERT** or **Transformers** for improved accuracy.
- Expand the dataset to cover a broader range of news articles.
- Add support for multilingual news detection.