Student Number, Firstname and Lastname: 202011204 Furkan Çoban

Report: CENG 493 – Homework 2

---

1. Objective

This assignment's objective is to use machine learning algorithms to categorize news stories as **REAL** or **FAKE**. **Naive Bayes** and **Logistic Regression** were the two classifiers we used. Performance metrics including **accuracy**, **precision**, **recall**, and **F1-score** were examined after the dataset was preprocessed and numerical features were assigned vectorize using the **TF-IDF** representation.

---

2. Description of the Dataset

This assignment's dataset includes the following:

- **Text**: The news article's body.
- **Label**: The news article's classification "REAL" or "FAKE."

The "title" and "Unnamed: 0" columns were eliminated because they had no bearing on the classification task.

---

3. Approach

1. **Text Preprocessing**:
    - All text was lowercased.
    - Removed stopwords, non alpha charachters, and punctuation.
    - The text was tokenized into words.
    - Normalized the text by using lemmatization with NLTK.
2. **Feature Extraction**:
    - The preprocessed text was transformed into numerical vectors using **TF-IDF Vectorizer**, which has a feature count of up to 5000.
3. **Data Splitting**:
    - Using stratification to maintain the label distribution, divide the dataset into **80% training data and 20% testing data.**
4. **Model Implementation**:
    - Two classifiers were trained:
        - The **Naive Bayes Classifier** is an effective probabilistic model for text classification.
        - **Logistic Regression**: An effective linear model for issues involving binary categorization.
5. **Evaluation Metrics**:
    - Used the following metrics to assess both models on the test set:
        - **Accuracy**: The percentage of samples that are correctly classified.
        - **Precision**: The percentage of positive observations that were accurately predicted.
        - **Recall**: The model's capacity to recognize every positive sample.
        - **F1-Score**: The precision and recall harmonic mean.
          The model's performance across expected and real labels was visualized using a **confusion matrix**.

---

4. Results

The following are the outcomes for both classifiers:

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.8761	0.8707	0.8831	0.8769
Logistic Regression	0.9148	0.8971	0.9368	0.9165

Confusion Matrices:

- **Naive Bayes**:
  - True Positives (FAKE correctly labelled): 559
  - True Negatives (REAL correctly labelled): 551
  - False Positives: 83
  - False Negatives: 74
- **Logistic Regression**:
  - True Positives (FAKE correctly labelled): 593
  - True Negatives (REAL correctly labelled): 566
  - False Positives: 68
  - False Negatives: 40

---

5. Conclusion and Analysis

- In every metric, **Logistic Regression** fared better than **Naive Bayes**:
  - **Accuracy** increased from **87.61% to 91.48%**.
  - Logistic Regression showed consistently superior **F1-Score, Precision,** and **Recall**.
  - Logistic Regression is the favored method for this challenge since it showed superior capacity to correctly categorize both "REAL" and "FAKE" news articles.
- The confusion matrix for Logistic Regression shows much less False Positives and False Negatives compared to Naive Bayes.

---

6. Utilized External Libraries

The list of outside resources and tools utilized for this assignment is provided below:

- **Pandas**: Analyzing and manipulating data.
- **Scikit-learn**: Evaluation metrics and models for machine learning.
- **NLTK**: Text preprocessing (lemmatization, tokenization, stopwords).
- **Matplotlib**: Confusion matrix visualization.
- **Seaborn**: Confusion matrix heatmap plotting.
