INTELLIGENT DOCUMENT CLASSIFIER

BY: SANYA UPPAL

SANYA UPPAL

JANUARY 12, 2025

CLOUD COUNSELAGE
Cloud Counselage Pvt.Ltd.

Model 1: Traditional Model

Model Used: Logistic Regression Model

❖ Preprocessing Steps:

1. **Data Exploration:**
    - The dataset is loaded using pandas.read_csv().
    - Exploration methods include:
        - df.info() to check data types.
        - df.isnull() to check for missing values.
        - df.shape for dimensions.
    - Categorical column Category is analyzed to find unique categories, counts, and percentages.
2. **Data Visualization:**
    - **Bar Plot:** Displays the count of each category.
    - **Pie Chart:** Shows the proportion of each category.
    - **Text Length Distribution:** Analyzes the text length variation.
    - **Word Cloud:** Visualizes the most frequent words in the text.
    - **Category-wise Average Text Length & Frequent Words by Category:** Highlights variations in text data across categories.
3. **Data Preprocessing:**
    - **Encoding:**
        - The Category column is encoded using both **Label Encoding** and **One-Hot Encoding**.
    - **Text Preprocessing:**
        - **Tokenization** breaks text into words.
        - **Stopwords Removal** filters out common words like "the" and "and".
        - **Lemmatization** reduces words to their base form (e.g., "running" becomes "run").
        - **TF-IDF Vectorization** converts text data into numerical features.
        - **Latent Semantic Analysis (LSA)** using **Truncated SVD** reduces dimensionality and uncovers hidden structures.

❖ Architecture and Methodology for Logistic Regression Model:

- **Model Building:**
    - Data is split into training and testing sets using train_test_split.
    - **Logistic Regression** is chosen to classify documents.
    - **Grid Search** is employed to tune hyperparameters for optimal performance.

SANYA UPPAL

- **Model Evaluation:**
  - Evaluated using **accuracy**, **precision**, **recall**, and **F1-score**.

MODEL 2: Deep Learning Model

Model Used: LSTM (Long Short-Term Memory) & Word2Vec

❖ **Preprocessing Steps:**

1. **Word2Vec Model:**
   - Converts words into vector embeddings based on their context using either **CBOW** or **Skip-gram** architectures.
   - The model is saved for future use after training.
   - **Embeddings Generation** converts words into vector representations.
2. **Data Splitting and Scaling:**
   - The dataset is split into training and testing sets.
   - **Scaling** is applied to ensure uniformity across features.

❖ **LSTM Model Architecture and Methodology:**

- **Dense Layer**: 512 units with **LeakyReLU** activation.
- **Regularization**: L2 regularization to avoid overfitting.
- **LSTM Layer**: The core network used in NLP tasks.

SANYA UPPAL

Evaluation Results and Comparison:

Logistic Regression Model:

- **Accuracy:** 99%
- **Precision:**
  - Blog: 0.98
  - E-commerce: 1.00
  - Legal: 1.00
  - News: 1.00
  - Scientific: 1.00
- **Recall:**
  - Blog: 1.00
  - E-commerce: 1.00
  - Legal: 1.00
  - News: 0.98

    - Scientific: 1.00
- **F1-Score:**
    - Blog: 0.99
    - E-commerce: 1.00
    - Legal: 1.00
    - News: 0.99
    - Scientific: 1.00
- **Overall Accuracy:** 99%
- **Macro Average:**
    - Precision: 1.00
    - Recall: 1.00
    - F1-Score: 1.00
- **Weighted Average:**
    - Precision: 1.00
    - Recall: 0.99
    - F1-Score: 1.00

LSTM Model:

- **Test Loss:** 1.072
- **Test Accuracy:** 50.8%

SANYA UPPAL

Conclusion:

- The **Logistic Regression Model** performs significantly better, with high accuracy, precision, recall, and F1-score across all categories.
- The **LSTM model** has a relatively low test accuracy of **50.8%**. Given that the data does not have a temporal or sequential nature, the LSTM model is unnecessary and over-complicated.
- The **Logistic Regression** model is more suitable for this task due to its superior performance and simplicity.

Challenges Faced and Solutions:

1. **Challenge 1:** Handling missing values and unbalanced categories.
    - **Solution:** Applied imputation techniques and oversampling methods like SMOTE for balancing categories.
2. **Challenge 2:** Text data preprocessing, particularly handling stopwords and lemmatization.
    - **Solution:** Used NLTK for efficient tokenization and stopwords removal, and leveraged pre-trained word embeddings for lemmatization.
3. **Challenge 3:** Model overfitting with deep learning models.

  - **Solution:** Applied regularization techniques such as L2 regularization and dropout to prevent overfitting in the LSTM model.

4. **Challenge 4:** Managing computational resources for deep learning models.

    - **Solution:** Chose Logistic Regression as the final model, considering its lower computational demands and higher performance.

Final Decision:

- After careful evaluation, the **Logistic Regression Model** is chosen for deployment due to its excellent performance on all metrics. The LSTM model, while powerful for sequence data, does not provide significant benefits for this task.

SANYA UPPAL