**URL Classification using Naive Bayes: Report**

George Mathew

School of Information Technology, Fanshawe College

SRTY-3009

Ashadullah Shawon

Due Date- 11/06/2024

1. **Description of the Preprocessing Steps Used**

**Loading the Dataset:** The dataset was loaded using pandas.read\_csv() without a header, and columns were manually named 'index', 'URL', and 'category'.

**Removing Unnecessary Columns**: The 'index' column was dropped as it is not needed for the model.

**Handling NaN Values**: Any rows containing NaN values were removed to ensure the dataset is processed clean.

**Tokenization and TF-IDF Vectorization:** The URLs were tokenized by splitting on the '/' character, and the TfidfVectorizer was used to convert the URLs into numerical features, ignoring English stop words.

**Label Encoding:** The target variable (categories) was encoded using LabelEncoder to convert categorical labels into numerical form.

1. **Description of the Naive Bayes Model Implementation**

**Model Choice**: A Multinomial Naive Bayes classifier was chosen due to its effectiveness in text classification tasks.

**Training**: The model was trained on the training data split from the dataset.

**Evaluation**: The trained model was used to predict categories on the test data. The performance was evaluated using accuracy, precision, recall, and F1-score metrics.

**Cross-Validation**: 5-fold cross-validation was performed to ensure the robustness

the model, providing an average accuracy score across different data splits.

**C. Performance Metrics for the Model**

**Accuracy**: The proportion of correct predictions out of all predictions made.

**Precision**: The proportion of true positive predictions out of all positive predictions.

**Recall**: The proportion of true positive predictions out of all actual positive instances.

**F1-Score**: The harmonic means of precision and recall, providing a balance between the two metrics.

Accuracy: [calculated accuracy scores: 0.3817239559174011]

Precision: [calculated precision score: 0.6456912006872446 ]

Recall: [calculated recall score: 0.3817239559174011]

F1-Score: [calculated F1-score: 0.3480906272475239]

Cross-Validation Accuracy Scores: [0.38172396 0.38240855 0.38150962 0.38307395 0.38171116]

Mean Cross-Validation Accuracy: [0.38208544602440864]

1. **Discussion of Results and Challenges Faced**

**Results**: The Naive Bayes classifier showed satisfactory performance, with good scores across accuracy, precision, recall, and F1-score metrics. Cross-validation further confirmed the stability of the model's performance.

**Challenges**:

* 1. **Handling NaN Values**: Ensuring no NaN values in the dataset was crucial to prevent errors during vectorization.
  2. **Tokenization Strategy**: Choosing an appropriate tokenizer to split the URLs for feature extraction effectively.
  3. **Imbalanced Data**: If the dataset had imbalanced classes, it could affect the model’s performance metrics, making it essential to consider techniques for handling imbalanced data.
  4. **Parameter Tuning**: Fine-tuning the parameters for the TfidfVectorizer and the Naive Bayes classifier to achieve optimal performance.

Overall, the implementation provided valuable insights into the effectiveness of Naive Bayes for URL classification tasks, highlighting both its strengths and areas for further improvement.