Document Classification Model for FinacPlus: Detailed Analysis

Introduction:

In addressing the document classification task for FinacPlus, a meticulous approach was adopted to develop a robust machine learning model capable of accurately categorizing documents across five distinct categories. This involved extracting text content from HTML files and subjecting it to thorough preprocessing before training and evaluating the model.

Methodology:

The methodology employed focused on leveraging key libraries and techniques to streamline the document classification process. BeautifulSoup, a renowned web scraping library, was utilized to extract text data from HTML files. Following this, special characters were meticulously removed, and newline characters were addressed to ensure data cleanliness.

To enhance semantic understanding, lemmatization was applied by tokenizing the text, lemmatizing the tokens, and recombining them into coherent sentences. This approach aimed to capture the nuanced semantics of the documents accurately.

Subsequently, the data was partitioned into training and test sets using the train_test_split library, setting the stage for model training and evaluation. A range of classification algorithms, including Logistic Regression, Naive Bayes MultinomialNB, SVM SVC, DecisionTreeClassifier, RandomForestClassifier, and GradientBoostingClassifier, were initialized for evaluation.

Evaluation:

Rigorous cross-validation with a fold of 5 was conducted to evaluate the performance of each model, focusing primarily on accuracy as the key metric. Notably, SVM SVC emerged as the standout performer, exhibiting the highest accuracy among the models assessed.

To corroborate the findings, AU ROC, a widely accepted method for assessing classification models, was employed. The evaluation reaffirmed the superior performance of SVM SVC, further validating its suitability for the document classification task.

Additionally, KFold validation was utilized to reinforce the conclusions. Across all validation methodologies, SVM SVC consistently demonstrated robust performance, culminating in an impressive accuracy score of 94%.

Conclusion:

In conclusion, Taking the accuracy as the key metric, SVM SVC is a preferred model for this classification.