[Document Classification Project Report: Resume vs. Non-Resume Documents.]()

By - Divyanshu Yadav

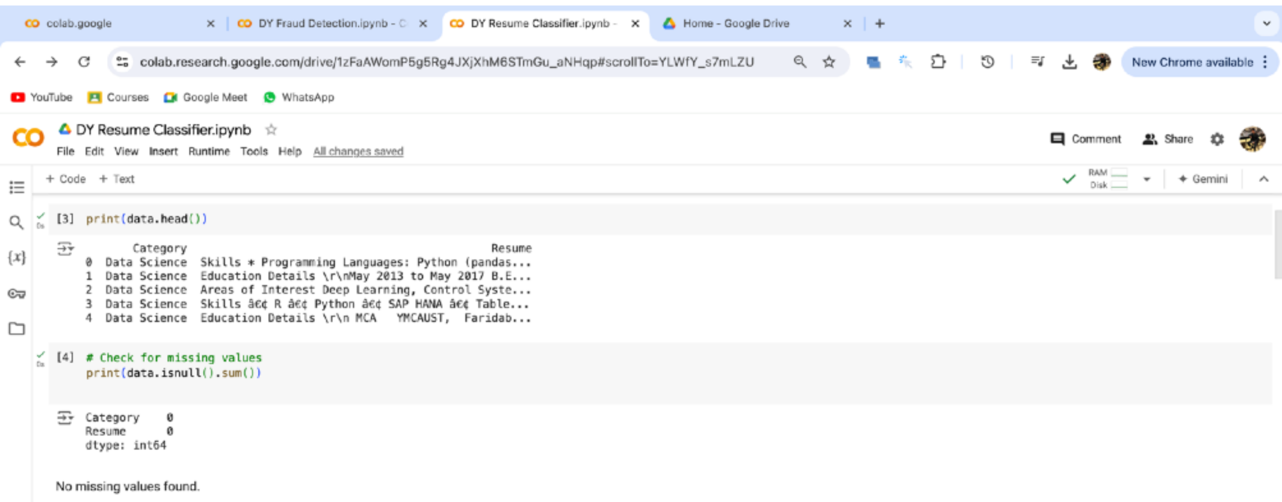
Link to Google Colab Notebook:- https://colab.research.google.com/drive/1zFaAWomP5g5Rg4JXjXhM6STmGu_aNHqp?usp=sharing

1. Introduction

- **Objective**: The aim of this project is to build a machine learning model that classifies text documents as resumes or non-resumes. This project focuses on preprocessing text data, feature extraction, and model training and evaluation.
- **Dataset Link**: https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset?resource=download
- **Approach Overview**: The project followed these major steps:
    - Data Loading and Exploration
    - Text Preprocessing
    - Feature Extraction using TF-IDF
    - Model Building with Logistic Regression
    - Model Evaluation with metrics like accuracy, AUC, and confusion matrix
    - Feature Importance Analysis

2. Data Loading and Exploration

- The dataset was loaded from a CSV file into a Pandas DataFrame.
- Explored the first few rows of the dataset to understand its structure and ensured there were no missing values.



3. Text Preprocessing

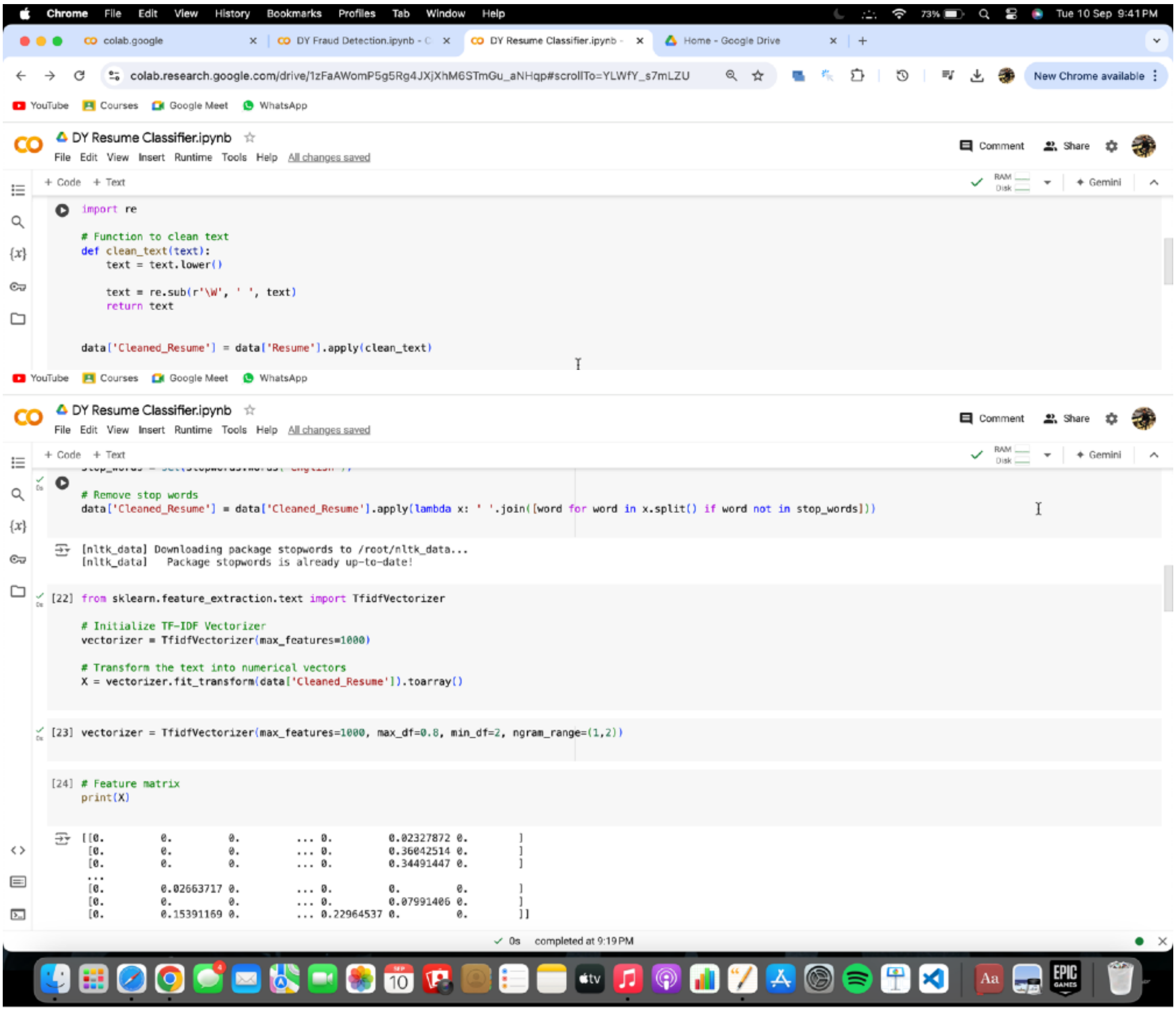
Challenges: Text preprocessing is crucial for preparing raw text data for machine learning models. The main challenges were handling inconsistencies in text formatting, stop words, and stemming/lemmatization.

Steps:

- **Cleaning**: Converted all text to lowercase and removed non-alphanumeric characters using regular expressions.
- **Tokenization**: Split the text into individual tokens using nltk.
- **Stop Words Removal**: Removed common words (like "the", "is", etc.) that do not provide significant value.
- **Lemmatization**: Reduced words to their base form to handle different word variations.

4. Feature Extraction using TF-IDF

Approach: To convert the text data into numerical features, I used the **TF-IDF (Term Frequency-Inverse Document Frequency)** method. I limited the maximum number of features to 1000 and also applied n-grams to capture patterns of two or more words.



5. Model Building

Algorithm Choice: I used **Logistic Regression**, a robust and interpretable classification algorithm, as the primary model for this task.

- The data was split into training and testing sets (70%-30% split).
- Cross-validation was used to validate the model across different subsets of data.
- Hyperparameter tuning was performed using gridsearchcv to optimize the model.

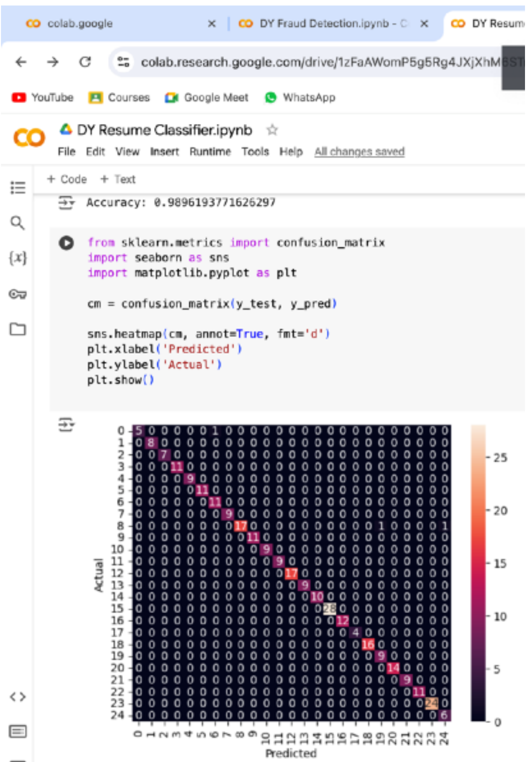
6. Model Evaluation

Metrics Used:

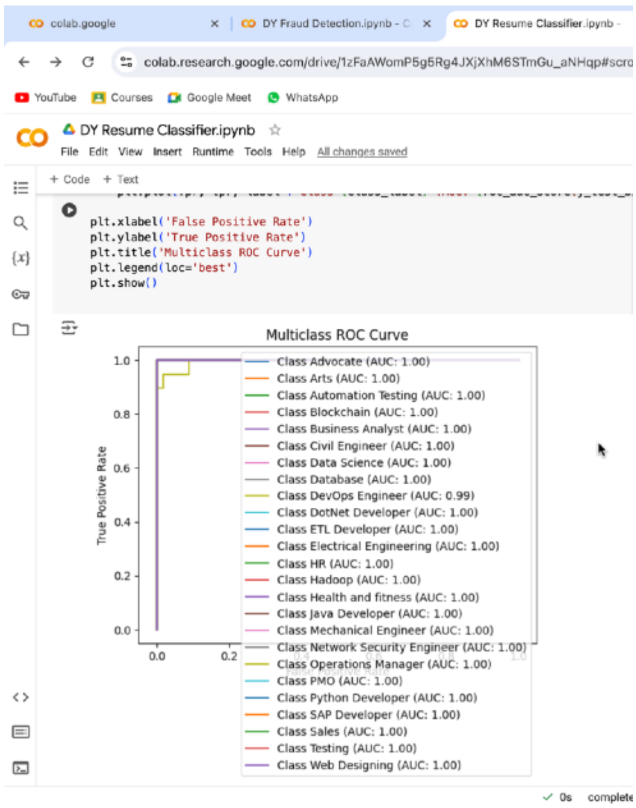
- **Accuracy**: The model achieved an accuracy of ~99%, indicating strong performance on the test set.
- **Confusion Matrix**: The confusion matrix was plotted to visualize the model's prediction performance.
- **ROC Curve and AUC**: ROC curves were plotted for each class, and the AUC scores were calculated to measure the model's classification performance.

Visualization:

- Confusion Matrix:



ROC Curve:



8. Challenges Faced

- **Text Preprocessing**: Handling inconsistencies in the data, such as formatting issues in text, required careful cleaning and tokenization.
- **Class Imbalance**: In some cases, certain categories had fewer samples, which could affect the model's ability to generalize. I addressed this using proper cross-validation.
- **Hyperparameter Tuning**: Fine-tuning the model's hyperparameters took several iterations to find the optimal values for performance

9. Conclusion

- The model successfully classified resumes with high accuracy and demonstrated strong performance across all metrics.
- Text preprocessing and feature extraction played a key role in improving model performance.
- Logistic regression proved to be effective for this task, though other models could be explored in future iterations.

Accuracy: The logistic regression model achieved an accuracy of **98.96%**.

AUC: The AUC score for multiclass classification using the 'ovr' strategy was **0.99977.**

Most Important Features: The words contributing the most to the classification were ['cases' 'com' 'maharashtra' 'details' 'board' 'accounts' 'january' 'university' 'mumbai' ‘legal']

By:- Divyanshu Yadav
divyanshuydv0002@gmail.com
+91 8588889331