Week 10: Deliverables

Group Name: NLP_Task_Force – Document Classification

Name: Konstantinos Soufleros

Email: soufleros.kostas@gmail.com

Country: Serbia

Company: Data Glacier

Specialization: NLP

Github: https://github.com/kostas696/DG_Intern/tree/main/week10

Internship Batch: LISUM30

Date: 09/04/2024

PROBLEM DESCRIPTION

The task involves analyzing a dataset comprising newsgroup documents categorized into various topics. The primary aim is to gain insights into the content of these documents, identify prevalent themes within each newsgroup, and discern any underlying patterns or trends in the data.

DATA UNDERSTANDING

The dataset consists of documents from 20 different newsgroups, covering a broad spectrum of topics such as sports, religion, politics, and technology. Each document is labeled with its corresponding newsgroup category, providing a structured format for analysis.

TYPE OF DATA

The data is structured and tabular, organized into a pandas DataFrame with six columns: 'Newsgroup', 'Original_Content', 'Content_prep', 'Content_spacy', 'Prep_Document_Length', and 'SpaCy_Document_Length'. The 'Newsgroup' column contains categorical labels indicating the category of each document, while the 'Original_Content' column contains the textual content of the documents. Additionally, the dataset includes preprocessed versions of the text ('Content_prep' and 'Content_spacy') along with document length information.

MODELING

Best Model Selection

The dataset was split into training, validation, and test sets. Two feature representations were utilized: TF-IDF from the original content and TF-IDF from SpaCy preprocessed text. Two models were evaluated: Naive Bayes and SVM.

Best Model based on F1-score and ROC-AUC: SVM with TF-IDF from SpaCy preprocessed text.

Hyperparameter Tuning

Hyperparameter optimization using RandomizedSearchCV was employed for hyperparameter tuning due to its efficiency compared to grid search. The search space included parameters for SVM models. The best parameters and score were identified through Randomized optimization.

Further evaluation of the best fine-tuned model was conducted on the test set to assess its performance.

The results and best model were saved using pickle for future reference.

CONCLUSION

SpaCy-based text preprocessing offers a more efficient and comprehensive approach compared to the original method. By leveraging advanced NLP capabilities, SpaCy improves the accuracy and consistency of preprocessing tasks, leading to better-quality textual data for subsequent analysis and modeling. The comparison highlights the benefits of adopting SpaCy for text preprocessing in data science workflows.