Name: Kenilbhai Sutariya
Company: Visable
Project: ML challenge for working student
28 Jan 2024

Project Title: German Search Query Classification with REST API Deployment

My Approach in this Project:

The project will follow a structured approach to problem-solving, including:

1. Data exploration and understanding
2. Feature engineering and selection
3. Model evaluation and selection
4. Model optimization
5. API development and deployment

Problem Statement:

We aim to create a robust machine learning model that effectively classifies German search queries into predetermined categories. Our model will have the ability to handle a diverse range of search queries, encompassing short phrases, slang, and even misspellings.

Dataset Description:

The dataset consists of a CSV file containing two columns:

text: Represents the search queries in German language (short phrases)

label: Represents the associated category for each search query

Data Preprocessing:

The data will be preprocessed to clean and prepare it for modeling.

The process would involve:

1. Handling missing values
2. Cleaning the text data to remove punctuation, special characters, and stop words

Here, I can use tokenization, stemming, and lemmatization techniques in order to make our model training smooth, however I did not use it due to the different algorithm selection.

Feature Extraction:

I will extract features from the preprocessed text data to represent the search queries in a numerical format that is compatible with machine learning models. This process will include TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, where words will be assigned weights based on how often they appear in the document and how uncommon they are across the entire dataset.

Model Training:

A machine learning model will be trained on the preprocessed and feature-extracted data to learn to classify search queries into the predefined categories. Suitable models for this task include:

1. Support Vector Machines (SVMs): Powerful classifiers that can handle high-dimensional data
2. Decision Trees: Tree-based classifiers that can capture non-linear relationships in the data

Here I could use both algorithms as these algorithms are very powerful and allows flexibility to large dataset. So, I chose to go with decision tree algorithm.

Model Evaluation:

The trained model is evaluated on a separate test dataset to assess its performance. Key metrics for evaluation includes Accuracy:

Accuracy: The proportion of correct classifications

The accuracy of model is always been above 80%.

FAST API Development:

A Fast API will be developed to allow users to classify search queries using the trained model. The API will consist of endpoints for:

- Sending search queries to the model for classification
- Receiving the predicted category for each query

Deployment:

The model and API will be deployed using Docker containers.

Future Work and Improvements:

1. Enhancing model performance through techniques like hyperparameter tuning and ensemble learning

2. Exploring more advanced language models for improved feature extraction and classification

3. Integrating the API with a search engine or application for real-world applications

Conclusion:

Through this project, I got opportunity to demonstrate my ability to develop and deploy a machine learning model for german search query classification, utilizing Fast API for practical application. I have showcased the process of data preprocessing, feature extraction, model training, evaluation, and deployment, providing a comprehensive understanding of the machine learning workflow.