Case Study: Multiclass Classification Model for Product Descriptions

Overview

Welcome to the Machine Learning internship case study at Insider. This case study aims to assess your skills in data science, machine learning, and software engineering. You will be working with a public dataset containing product descriptions and their respective categories. Your task is to build a multiclass classification model to predict product categories based on their descriptions.

Objectives

1. **Data Understanding and Preprocessing:**
    - Explore the dataset to understand its structure and characteristics.
    - Handle any missing or inconsistent data.
    - Perform necessary preprocessing steps such as tokenization, stemming/lemmatization, and vectorization of text data.
2. **Model Building:**
    - Experiment with different machine learning algorithms suitable for multiclass classification.
    - Implement hyperparameter tuning to optimize model performance.
    - Evaluate the model using appropriate metrics.
3. **Documentation and Reporting:**
    - Document your code and methodology clearly in a Jupyter notebook.
    - Include comments and markdown cells to explain your thought process and decisions.
    - Provide a detailed report on your findings, including data exploration, model selection, and evaluation.
4. **Model Deployment:**
    - Package your entire work in a containerized environment for easy deployment and reproducibility.
    - Ensure that the container can be run on any standard environment with minimal setup.
5. **Inference Interface:**
    - Create an interface that takes a product description as input and returns the predicted category.
    - Ensure that the interface is user-friendly and can be easily tested.

Guideline

1. **Dataset and Tools:**
    - Use the provided public dataset for product descriptions and categories.
    - You are free to use any libraries and frameworks you are comfortable with.
    - Ensure that all dependencies are listed and managed appropriately.
2. **Code Quality:**
    - Write clean, readable, and well-documented code.
    - Follow best practices for coding and version control.

    - Use comments to explain complex logic and document the purpose of different code sections.
3. **Model Evaluation:**
    - Use appropriate evaluation metrics for the mentioned problem.
    - Provide a thorough analysis of the model's performance, including any trade-offs and challenges faced.
4. **Containerization:**
    - Package your application and associated files into a container.
    - Ensure that the container includes all necessary dependencies and can be executed in a standard environment.
5. **Inference Interface:**
    - Develop a simple interface that allows for easy testing of the model's predictions.
    - Ensure that the interface is intuitive and provides accurate predictions based on the input description.
6. **Submission:**
    - Submit your Jupyter notebook, container configuration files, and any other necessary files.
    - Ensure that your submission includes clear instructions on how to set up and run the container, as well as how to use the inference interface.
    - Please upload your work to GitHub. We will evaluate your submission based on your GitHub repository.

Restrictions: You are not allowed to use LLMs when developing models (e.g ChatGPT)

Expected Deliverables

1. **Jupyter Notebook:**
    - Detailed exploration and preprocessing of the dataset.
    - Implementation of different models and their evaluations.
    - Comprehensive documentation and comments.
        i. Summary of your approach, findings, and insights.
        ii. Analysis of model performance and discussion of results.
        iii. Any additional observations or recommendations.
2. **Inference Interface:**
    - Productionize your application with a containerization method.
    - A simple, user-friendly interface for making predictions.
    - Documentation on how to use the interface.

Dataset Information

- Public Turkish e-commerce data with product id, description and category.
- 1800 products, all files contain first the ID of the product and separated from the value by ';' character.
- Product_Categories: Contains raw categories of the products that are in that dataset. It may not be available. Each category is separated by '>' character.
- Product_Explanation: Contains raw extra information of the products that are in that dataset. It may not be available.
- Datasets can be accessed via this Drive link.

ecem.alpagul@useinsider.com
umut.