abwab.ai

HomeWork

Background

Abwab.ai is an AI underwriting startup aiming to automate lending for Small and Medium-sized Enterprises (SMEs). As such, we have a lot of data to analyze and label for each SME which we need to filter to find the most relevant samples to feed our models or to guide the annotation team more efficiently towards important samples.

As part of our hiring process for a Senior Machine Learning Engineer, we'd like you to complete the following task by implementing outlier detection and distribution shift scoring methods. Finally, your method should be served by a simple API.

Task Description

1. Dataset Selection

You will be working with the following text dataset:

Amazon Reviews Dataset: This dataset contains multiple product categories. Please choose a subset of categories that will fit on your hardware since the dataset might be too big for your machine. Explain your choice of categories.

This dataset contains text reviews along with additional features such as ratings, timestamps, or user information. You'll find more details in the dataset page. Please use as many side features as possible for your models on top of using the text feature. Explain your choice of features.

2. Outlier Detection

Implement methods to detect outliers within the dataset. You should:

- Research and choose one or more appropriate outlier detection techniques.
- Test your approaches with appropriate metrics. Explain your metrics choice.
- Provide a clear explanation of your chosen methods and their implementation.

- Identify and report the samples most likely to be outliers.

3. Distribution Shift Scoring

To complement the outliers detection above, each sample detected as an outlier should have a corresponding interpretable, bounded and comparable score that measures the distribution shift of each sample. Your solution should:

- Implement methods to quantify how much each sample deviates from the training distribution.
- Ensure that the test set has a higher average drift score compared to the training set.
- Provide visualizations and explanations of your distribution shift scoring results.

4. Model Serving

Create a model serving solution that exposes your outlier detection and distribution shift scoring methods through an API. Your solution should:

- Use Docker or even model serving frameworks.
- Provide clear documentation on how to build and run your service.
- Include API endpoints for outlier detection and distribution shift scoring.
- Handle input validation and error cases appropriately.

5. Documentation and Code Quality

Your submission should include:

- A well-organized GitHub repository containing all your code.
- A comprehensive README file explaining:
    - Your approach and methodologies
    - How to set up and run your code
    - The structure of your repository
    - Any assumptions or limitations of your implementation
- Clear and concise code comments.
- Proper error handling and logging.

Evaluation Criteria

Your submission will be evaluated based on:

1. Understanding and implementation of outlier detection techniques
    a. Did you use simple methods? Or did you investigate more recent methods and papers?

2. Effectiveness of distribution shift scoring methods
3. Quality and clarity of your code and documentation
4. Proper use of software engineering best practices
5. Creativity and innovation in your approach
6. Scalability and efficiency of your solutions

Submission Instructions

1. Create a private GitHub repository for your project.
2. Include all necessary documentation in your repository.
3. Once completed, share the repository with these GitHub users: cpcdoy and basseko.
4. Please notify us by email once you're done with the task.

Next Step

When your task is completed, send it to us. We'll then schedule a call to discuss your approach and findings together, similarly to how we'd do it if we were already working together.

Time Frame

You will have **one week** from the date you receive this task to complete and submit your solution. If you need more time, please let us know, and we'll be happy to accommodate your schedule.

Good luck! We look forward to seeing your innovative solutions to this challenge.

If you have any questions or even if you want to discuss approaches together on your task, feel free to reach out to us, we'll be happy to discuss together since day to day this is how we'll be working together!