Modelización de problemas de la Empresa

Problem proposal

Title: Using LLMs for sentiment analysis

Description:

Large Language Models (LLMs) have revolutionized Natural Language Processing with their human-like text generation.

One of the key challenges when using LLMs is data privacy. Certain sensitive data, such as personal details, interviews, or medical records, cannot leave a company's servers under any circumstances.

In this project, we propose using an open-source LLM (Llama2) locally for feature engineering and subsequent sentiment analysis prediction on restaurant reviews, exemplifying a common industry pipeline when working with LLMs.

Development details:

You can download the data from https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews.

Then:

1. Download Llama-2-7b-chat (https://huggingface.co/TheBloke/Llama-2-7b-Chat-GGUF) onto your computer. Use it to perform sentiment analysis on the “Review” column, instructing the LLM to classify each review as positive, negative, or neutral. If available, leverage a GPU for inference. If processing the entire dataset is time-consuming on your machine, limit your analysis to 100 randomly selected rows.
2. With the newly extracted sentiment feature, alongside the number of reviews and the number of followers (which can both be extracted from the “Metadata” column), train a decision tree with the “Rating” column as the target. Ensure the tree doesn't have too many splits.
3. Visualize the decision tree. Identify and report the most impactful variable used by the tree.

You will be scored based on (in order of importance):

1. Your code cleanup, comments, code reproducibility, and using good programming practices.
2. A clear visualization of the model.

Please note that accuracy or performance will not be the primary scoring criteria. Instead, we are looking for clear and well-documented code, an understanding of the problem, and an interpretable model used to solve the problem.