Project Assignment: NLP/LLM Application in Stock Price Prediction

Project Overview

The goal of this project is to explore the application of Natural Language Processing (NLP) and Large Language Models (LLMs) in analyzing textual data for predicting stock price movements. Specifically, the study will focus on using financial filings (10-K, 10-Q, and 8-K) from the SEC's EDGAR system to extract signals for forward price prediction.

Data

- **Source**: SEC's EDGAR system.
- **Data Types**: You may focus on any of the following filing types: 10-K, 10-Q, 8-K, or all of them.
- **Download Instructions**: Refer to EDGAR Tools Documentation for guidance on downloading filing data.
- **Stock Price Data**: Use **Yahoo Finance** to query historical stock prices for the prediction target period. Yahoo Finance provides APIs and tools to retrieve stock price data efficiently. Ensure proper alignment of the stock price data with the filing release dates.

Study Scope

- **Period**: Focus on the full period or a sub-period between **2014–2024**.
- **Stock Universe**: Analyze stocks from either the **S&P 500** or **Russell 1000** index.

Prediction Target

- **Primary Target**: Predict the **stock return** from immediately after the report is released to **5 trading days later**.

Reference Material

- **Survey Paper**: Review the paper *"Generative AI and Finance"* by Andrea L. Eisfeldt & Gregor Schubert for insights into the applications of generative AI in finance.
- **Key Section**: Start with **Table II** in the paper to explore relevant application areas.

Application Areas

You may freely explore the following strategies or propose new ones for extracting forward price prediction signals using EDGAR data:

1. **Embeddings**
2. **Text Classification**
3. **Retrieval-Augmented Generation (RAG)**
4. **Hypothesis Generation**

Recommended Discussion Topics

While not mandatory, you are encouraged to discuss strategies related to the following questions:

1. Training Targets

- Besides stock price movement after the report release, what other prediction targets can you construct?

2. Fine-Tuning/Prompt Engineering

- For the language model you use, what fine-tuning or prompt-engineering strategies will you employ to enable the model to learn effectively from textual data?

3. Avoiding Data Look-Ahead

- How will you address the issue of **data look-ahead**?

Deliverables

- A detailed report.
- Code implementation.

Resources

- EDGAR Tools Documentation
- Yahoo Finance API Documentation**:** Use Yahoo Finance to query historical stock prices and ensure proper synchronization with EDGAR filing release dates.
- *"Generative AI and Finance"* by Andrea L. Eisfeldt & Gregor Schubert

Good luck with your project!

Please keep this project confidential, your submissions will also be kept confidential.