Project Proposal – CS7290

Sashank Vasepalli

Introduction:

Large Language Models are taking the world by storm. They have crept into every industry, looking to give a significant technological lead to its early adaptors. The most popular and in-demand jobs since Q2 2023 are in Generative AI and LLMs. However, one of the industries that is slow to adapt is Finance. This stems from the combination of the risk aversion nature of the financial sector, insufficient infrastructure, data privacy concerns, and a lack of interpretability.

Objective:

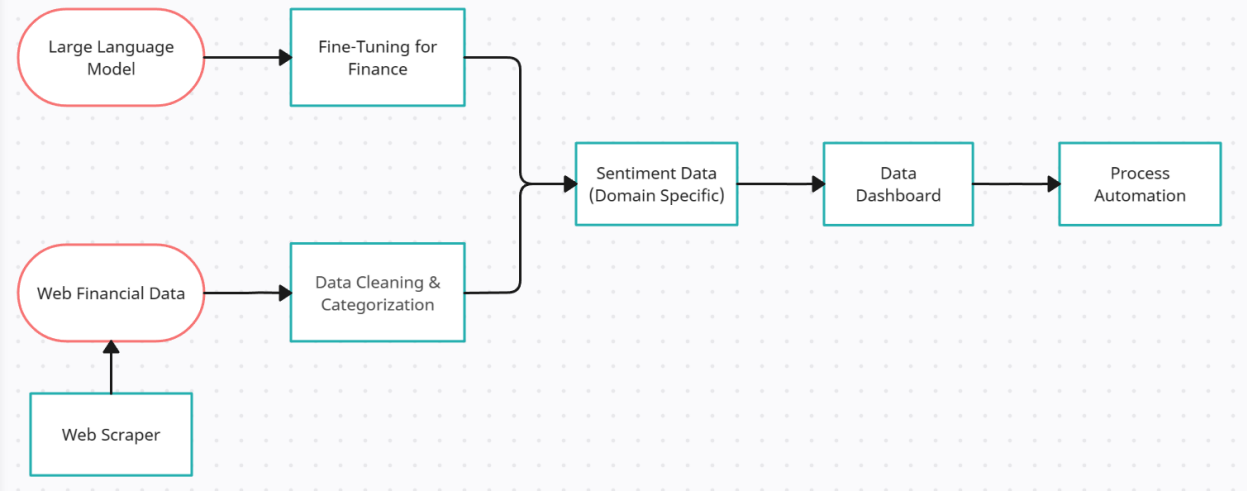
My objective for this project is to develop a system that performs sentiment analysis on financial data that is readily available on the internet, using LLMs. Due to the lack of high quality, open-source fine-tuned LLMs for finance [1)], I aim to build a tool to provide up-to-date sentiment data for different financial domains.

Scope:

The objective of my system is to fine-tune a LLM or/and use a LLM pre-tuned for finance [2)] for sentiment data for different financial domains across varying time frames: Short (10 Day), Medium (50 Day), Long (200 Day). I plan to develop the web scraper for daily live financial data, clean and extract useful metadata, use an LLM for sentiment analysis across financial domains, and provide an online dashboard. As a stretch objective, I plan to automate the process and host it using leading cloud services.

I do NOT plan on developing a trading algorithm using sentiment, as trading algorithms by themselves are extremely complex, and high-quality sentiment data would only serve to be a small part of the ensemble trading system [3)].

Methodology:



1. Select an LLM such as LLaMA 7B [4)], which can be fine-tuned.
2. Find relevant datasets and fine-tune the Large Language Model using techniques like QLoRA [5)].
3. Identify widely used financial documents on the web, ideally with APIs or readily available web scrapers (Ex. Reddit-WallStreetBets, Wall Street Journal, Twitter, Yahoo Finance).
4. Develop a web scraper for scraping and storing necessary data using Python and Beautiful Soup/Selenium/Scrapy.
5. Clean incoming scraped data, extract metadata (views, outreach), and categorize into the following different domains: Equity Markets, Fixed Income, Commodities, Forex, Real Estate, Private Equity, Cryptocurrency.
6. Feed data to the LLM using good Prompt Engineering, to obtain sentiment data and confidence scores.
7. Devise an algorithm to add sentiment data and confidence scores, to generate sentiment across varying time frames: Short (10 Day), Medium (50 Day), Long (200 Day).
8. Create an easily digestible data dashboard to display the above data.
9. Stretch: Automate processes and host on online using cloud services.

References:

1) Large Language Models in Finance: A Survey (arxiv.org)

2) AdaptLLM/finance-LLM-13B · Hugging Face

3) Sentiment Trading and Hedge Fund Returns - CHEN - 2021 - The Journal of Finance - Wiley Online Library

4) meta-llama/Llama-2-7b-hf · Hugging Face

5) oobabooga/text-generation-webui: A Gradio web UI for Large Language Models. Supports transformers, GPTQ, AWQ, EXL2, llama.cpp (GGUF), Llama models. (github.com)