Project Scoping

Team 3 – Amber Lee, Cindy Zhang, Hejia Zhang, Jingwen Li, Wafer Hsu

I. Problem

1. Overview of an industry, business or problem

Over the years, investment firms have increasingly embraced technology and data science to forecast financial market trends. In today's era of abundant data availability, investors have the opportunity to make well-informed decisions by leveraging historical trends and real-time news information. However, the event-driven price changes are still an unpredictable part to all investors.

Our project aims to enhance stock price prediction by employing a comprehensive approach that integrates news content analysis, historical stock returns, and idiosyncratic exposure to Fama-French risk premiums. By combining these factors into a predictive model, we aim to harness the power of ubiquitous data to empower investors of all scales in making more accurate and effective investment decisions.

2. Why this is interesting and important

The utilization of natural language processing techniques to analyze news data for predicting stock market movements presents a novel avenue to advance research in deciphering the predictive power of news. While traditional financial analysts heavily rely on quantitative trends, patterns, and fundamental information, these factors may provide valuable insights into estimating and quantifying the systematic risk associated with equities. However, they often fall short in capturing the nuanced idiosyncratic or non-systematic risk, including sudden price changes, which cannot be fully reflected through conventional numerical indicators.

By employing advanced sentiment analysis, textual parsing, and machine learning algorithms, the analysis of financial news data can precisely identify sentiment shifts, extract relevant information, and uncover hidden patterns that may have a significant impact on stock price fluctuations. This integration of news analytics has the potential to enhance the accuracy and timeliness of stock market predictions, providing investors with a competitive edge in decision-making.

The successful application of news data analysis in stock market prediction could potentially revolutionize the investment landscape, empowering both institutional and retail investors to make more informed and proactive investment decisions. The resulting economic impact would extend beyond individual portfolios, influencing market dynamics and contributing to the overall efficiency and stability of financial markets.

In conclusion, leveraging cutting-edge techniques in natural language processing and machine learning to analyze financial news data opens up new possibilities for predicting stock market movements with greater precision and accuracy. By embracing this innovative approach, we can unlock valuable insights, transform decision-making processes, and shape the future of quantitative finance.

II. Proposal

1. Define the specific problem that should be solved

The problem solving process can be divided into two parts: text data sentiment analysis and time-series data prediction. The following graph shows the basic steps and models to be used in this project:

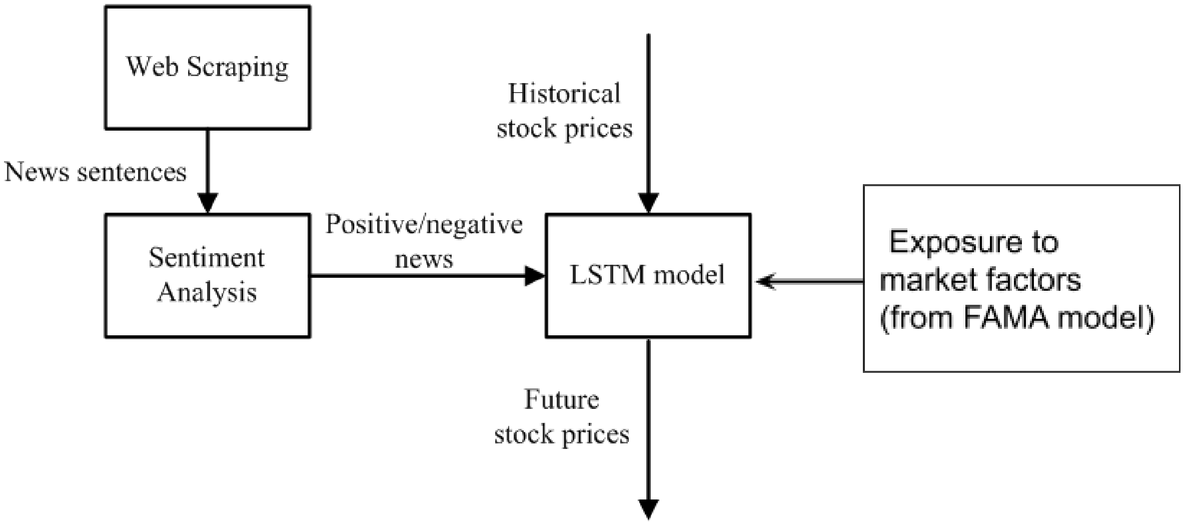

Fig 1 Basic schematic graph of the proposed problem

Our proposal covers diverse types of models, which is helpful for our team members to gain various aspects of skills throughout the project. The workflow also includes enough flexibility, which allows our team to work on different models simultaneously. It would be useful to compare our final predictive model with a baseline stock price predictive model to discuss the impact of news on the conventional stock price prediction performance.

III. Data

1. Universe:

The dataset for our project encompasses approximately five years of data, starting from January 1, 2018. To ensure sufficient liquidity and availability of news and price information, we have selected the most recent stocks from the S&P 500 index that were listed prior to January 1, 2018, as our universe of stocks.

2. Stock Prices:

The stock price data utilized in this project is sourced from Quandl via the NASDAQ data link. To mitigate the impact of stock splits and dividends, we have chosen the daily adjusted closing prices as our primary focus for predicting values.

3. Financial News:

The headlines of financial news articles relevant to the equities within our defined universe are extracted using the BeautifulSoup package. This extraction process is performed on news articles retrieved from Yahoo Finance.

4. Fama-French Risk Premiums Exposure:

To assess the exposure of each stock in our universe to the Fama-French risk premiums, we conduct linear regression analysis. This analysis involves regressing each stock's daily returns against the corresponding daily Fama-French factor returns.

The concept of consistent exposure to risk premiums, as emphasized by the Fama-French model, is taken into consideration. To maintain this concept, we employ a rolling linear regression approach over a 100-trading-day period. This allows us to capture the evolving exposures to each risk premium as our independent variables. We believe that these predictors effectively capture the expected changes in stock prices, while the sentiment analysis of financial news focuses on capturing idiosyncratic risks.

By utilizing these data sources and analytical techniques, our project aims to provide comprehensive insights into stock price prediction, leveraging both quantitative factors such as Fama-French risk premiums and qualitative factors derived from financial news sentiment analysis.

IV. Methodology

1. Time series data prediction

The long short term memory neural network (LSTM), which is good at learning order dependence in sequence prediction problems, will be applied to forecast the future stock price based on the previous prices. Historical time-series data (for example, 60-days historical prices) will be used as our feature data to predict the following days' time-series trends.

2. Web scraping

We will leverage the powerful capabilities of the BeautifulSoup package, a Python library for web scraping, to extract financial news headlines from Yahoo Finance. This package allows us to navigate and parse the HTML structure of the web pages, enabling efficient extraction of relevant textual information.

3. Sentiment analysis

Once we have obtained the financial news headlines, we will apply advanced Natural Language Processing (NLP) techniques to perform sentiment analysis. NLP involves using algorithms and statistical models to understand and interpret human language. In our case, we will use NLP to analyze the sentiment expressed in the financial news articles. By determining whether the sentiment is positive, negative, or neutral, we can gain insights into market sentiment and investor sentiment towards specific stocks or financial events.

The sentiment analysis results will serve as new variables in our predictive model for the stock market. Combined with other return factors, such as historical stock prices, trading volumes, and financial indicators, we will build a comprehensive model to forecast stock market movements. By incorporating sentiment analysis, we aim to capture the impact of market sentiment on stock price fluctuations and enhance the accuracy of our predictions.

4. Company's exposure to market factors

The Fama-French five factor model is used to estimate excess return of an investment asset. The factors in this model are 1) market factor (MKT) which is excess market return; 2) size factor (SMB) which is excess return of small versus large companies; and 3) value factor (HML) which is excess return of high book/market versus low book/market companies (value stocks versus growth stocks); 4) profitability factor (RMW) which is the difference between returns of firms with high and low profitability; and 5) investment factor (CMA) which is the difference in returns between high and low investment firms.

From this post, the model is given as

$$R_{ft} - R_{Ft} = \beta_0 + \beta_1 MKT + \beta_2 SMB + \beta_3 HML + \beta_4 RMW + \beta_5 CMA + \varepsilon$$

where the dependent variable $R_{ft} - R_{Ft}$ is the difference between the return in a period t and risk free rate and the five factors are described as above. We will estimate the coefficients with OLS for each company and time period t, so we will have $\beta_{i,\, company,\, t}$ to convey the exposure of the company to each market factor at period t. This results in another time series that will be used as input in the LSTM.

We have a strong conviction in the significance of assessing the exposure of each equity to the Fama-French risk premium as a means to analyze and predict expected returns. To capture the evolving nature of these exposures, we employ the expected exposure from the preceding 100 trading days as a reliable estimate for the subsequent day's exposure. This approach allows us to effectively account for the systematic risk inherent in stock prices.

By combining this quantitative analysis with the sentiment analysis of financial news, which primarily emphasizes idiosyncratic event-driven risks, we strive to provide a comprehensive assessment of the total risks associated with each equity. This holistic approach ensures that both systematic and idiosyncratic factors are adequately incorporated, allowing us to offer a well-rounded evaluation of equity risks and facilitate more informed investment decisions.

V. Summary

Overall, this data science project will harness the capabilities of web scraping, NLP, Fama-French model, and predictive modeling to provide valuable insights into the stock market. Financial news sentiment analysis as a crucial component of our predictive framework. By including the coefficients Fama-French model, we also account for broader macroeconomic conditions.