Assignment 2

CPSC 488: NATURAL LANGUAGE PROCESSING
10/14/2024
DR. CHRISTOPHER RYU
CALIFORNIA STATE UNIVERSITY, FULLERTON (CSUF)

Assignment Information

This assignment can be completed individually or by a team with up to 5 members.
Total score: 100
Due date: 12/01/2024 11:59PM
Members: Hammad Sheikh, Matthew Do, Ryan Avancena
Work distribution: Equal

Objectives

In this exercise, you will learn how to develop a sentiment analysis model and an automated trading system relying on the analysis results. Only Python programs written using Python 3.0 or higher will be accepted. NO Jupyter notebook or any other Python variants will be accepted for efficient grading.
Change: get news articles for at least 4 weeks for all the members of ETF FNGU

About the dataset and problem to solve

A stock, in investment terms, represents ownership in a company. When you buy stocks, you essentially buy a small piece of that company. The stock price for a company is mainly determined by the current value of the company that can be computed by the company's past financial data (e.g., earnings, revenues, profits, etc.) and future value determined by the company's future performance, which is unknown. The intrinsic value of a company refers to the actual and inherent worth of the company based on both current and future value. Therefore, a stock price is determined by the current value and estimated (perceived) future value of the company. Investors buy stocks based on the company's estimated intrinsic value. That's why the stock prices inherently fluctuate depending on the market situation and the company's future prospectus, which are typically published in news articles. The more information investors have, the better they can estimate the intrinsic value.
An Exchange-Traded Fund (**ETF**) is a type of investment fund that holds a collection of assets such as stocks, bonds, commodities, or a mix of these. There are many ETFs covering specific sections of the market currently being traded in the market. Some ETFs are conservative (slow price change) and some are highly aggressive or volatile in price change. **FNGU** is one of the ETF funds that concentrates on the top 10 popular technology companies, also known as **FANG+**, which include META, AAPL, AMZN, NFLX, NVDA, GOOGL, MSFT, CRWD, AVGO, and NOW. The fund manager occasionally changes the holdings. FNGU is highly volatile due to high leverage (3x). The price of FNGU change when the price of any of these holdings change. As a pair of FNGU, **FNGD** follows the inverse of FNGU for those who want to bet against FNGU. If most of the holdings are up, FNGU will go up 3x, but FNGD will go down 3x, and vice versa. In other words, the price movement of FNGU and FNGD is inverse, so you can buy and sell either one for profit in either direction as long as you can correctly predict the direction without relying on a sell-short strategy. Due to the popularity, the size, and their technological innovations of all those companies in FNGU, they most likely generate lots of news articles that impact the price of FNGU as well as the market indices.
In this assignment, you will analyze the FNGU price data and related news articles to develop a simple trading system. Your trading system will be equipped with a sentiment analysis model that analyzes all the (directly or indirectly) FNGU-related news articles for at least four weeks, computes its market impact, and trades only FNGU, FNGD, or both FNGU and FNGD based on the impact. For example, your trading system will buy a certain number of shares if the news article is considered positive and the balance is enough to cover the trade. Otherwise, the system will sell a certain number of shares if the news article is negative and a sufficient number of shares are already owned. The ultimate goal is maximizing the return from the initial investment, utilizing the sentiment analysis model.

Required activities

Utilize the **GPU** resources available on the **Nautilus** through the **Kubernetes** for modeling (**NOT** by the JupyterHub) and **write** an analysis report about your system's modeling results and trading performance by answering all the questions below with your **justification supported by the data**.

1. **Download** the historical price data for FNGU (or FNGD or both) and its (directly or indirectly) related news articles for your selected period of four

weeks from any free data sources (e.g., Yahoo Finance) and save the data in a JSON file format on your local machine. JSON is an open standard file format used for data storage or exchange. The price data should include market date, open price, high price, low price, closing price, and volume. The news articles may be related to the entire market, not necessarily only to FANG+. The direct FANG+-related news can be articles about the companies that you can download from the data source. Examples of indirect FANG+-related news can be the news about the market (Dow Jones ^DJI, S&P ^GSPC, Nasdaq ^IXIC, or the industry FANG+ belongs to). Briefly describe the types of news articles you downloaded for your system, explaining why and how you downloaded the data.

2. **Develop** a sentiment analysis model using Multilayer Perceptron (MLP) that can quantitatively estimate the impact of each news article on the FNGU price. Briefly describe

   (a) The methods used to (pre)process the news and price data for your MLP algorithm with justification and

   (b) the method(s) used to analyze the news and quantify its impact that will be used for trade.

3. **Backtest** your trading system based on the sentiment analysis model developed in (2) and measure its performance. Backtesting means testing the effectiveness of trading systems that utilize specific trading algorithms. In this case, your sentiment analysis model is a trading algorithm. For the backtesting, assume your account has an initial investment balance of $100,000, and buy/sell orders will always be filled without trading fees.

   To evaluate the performance of your trading algorithm, you need to **develop a simple trading system** that will buy a certain number of TSLA shares only if the balance is sufficient to cover the purchase and the positive market impact computed by the model or sell a certain number of shares only if it already holds enough number of shares (no short sell allowed) and the negative market impact. **Evaluate your system's trading performance** so far by calculating the following simple metrics

   (a) $gain or $loss for each trade,

   (b) the total *gainor*loss for all trades,

   (c) % return compared to the initial balance.

   Log every trade, including the key transaction data such as

   (a) the transaction date,

   (b) trading type buy/sell,

   (c) # of shares traded,

   (d) $amount used for the trade, and

   (e) the current balance after the trade to a log file "trade_log.json" for future analysis, verification, or accounting purpose.

   Display the trading summary, including the

   (a) total $gain or $loss for all trades and

   (b) % return compared to the initial balance ($100,000.00).

4. **Briefly describe** at least two methods or techniques (based on the relevant topics discussed in class) to improve the model performance and evaluation results on whether or not those methods improved the trading performance.

5. **Create word embeddings** based on Word2vec, other embedding method, or pre-trained embedding for your model and compare the trading performance with the best model without relying on word embeddings.

Warning: Although you can reuse any source codes available on the Internet, you are not allowed to share your codes with any other team or students in this class. Any student or team violating this policy will receive a **ZERO** score for this assignment, potentially for all the remaining assignments.

What to submit

- One analysis report includes all your member names, % contribution made by each member, and all the answers to the questions based on your analysis in **PDF** or **Word format**. If every member contributed equally, simply state "equal contribution." If your team does not agree on individual contributions, briefly write a task description for each member. Different grades may be assigned based on individual contributions, even if a group completed the work.

- **Upload one analysis report file** and **Python program file(s)**, individually. Please DO NOT upload any zip file since Canvas cannot open it.

- When you show some example data in your analysis report (when necessary), select only a few examples, not including the entire dataset.

- **Submit only one for each team.**

Grading criteria

- The overall quality of work shown in the report about modeling results, supporting data, analysis process, methods used, and correctly implemented programs
- The level of understanding as reflected in the report
- Effort (10%)

Abstract - Machine learning (ML) and Artificial Intelligence (AI) worlds are experiences tremendous growth. Their applications are uncountable. One realm of applications for ML and AI is Natural Language Processing (NLP). In this aspect, there are various classification models that assist with text classification and processing. Our goal for this assignment is to utilize NLP techniques in analyzing news data for ETF FNGU to build a ML model that would maximize our investment returns. The stock data can be obtained via the Python libraries that Dr. Ryu [1] defined. News articles will be scraped from the web.
Keywords - Stock Trading, ETF, FNGU, Natural Language Processing, Machine Learning

Contents

1 Introduction

There is a lot of stock data available in the world. All the stocks are impacted by the environment and the economical status of the world. This information is shared via news articles, which are in text format. We need to able to analyze news data, identify patterns and sentiment, and utilize it to predict impact on stocks. However, due to the sheer size of data available and the amount of convolution, it is not an easy task for a human. If we can utilize the computing power and technology available in computers, we may be able to make an impact. This may not have been possible a few decades ago. However, with the expansion of ML and AI technologies, numerous applications become available. With the use of NLP, we can perform text classification and sentiment analysis.
Without generalizing too much, let us look at an ETF, ticker FNGU, with the goal of applying existing text classification and sentiment analysis algorithms on news and stock data, and analyzing their results for performance and accuracy to maximize our investment returns.

2 Approach

The generalized approach is shown below in Figure 1. (Data Acquisition → Data Pre-Processing → Data Classification and Sentiment Analysis → Model Development for Investment ↔ Parameter and Feature Adjustment to Maximize Investment Returns.) Further details will be discussed in the following sections.

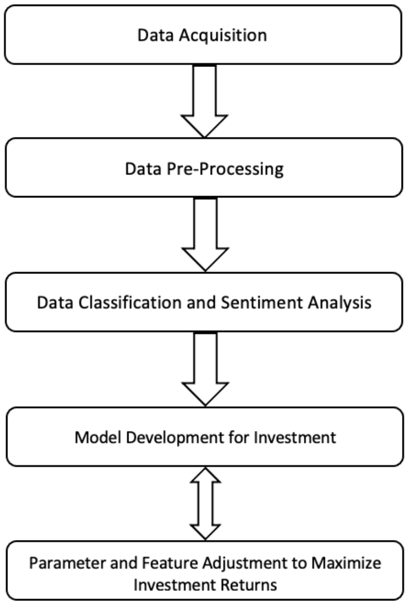

Data Acquisition
Data Pre-Processing
Data Classification and Sentiment Analysis
Model Development for Investment
Parameter and Feature Adjustment to Maximize Investment Returns

Figure 1: Generalized Approach

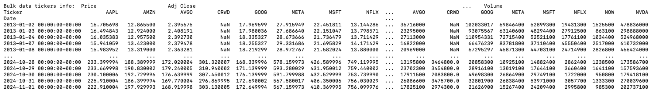
We will be relying on existing libraries as much as possible. This will help to focus on the classification and sentiment analysis task instead of building code from scratch for things that already exist. For example, train_test_split from sklearn.model_selection [2] will be utilized to split the data into training and testing datasets.

2.1 Data Acquisition

The first step in data processing, regardless of whether it is for our assignment or any other project, is data acquisition. We need to have a dataset that can be used to train need based models.
Our assignment needs two main datasets; one for stock information and one for news articles. For the ETF information, we utilized Python library yfinance [3] to

download the data and stored it locally in JSON files. For the news articles, we utilized get_news() function of the yfinance library [3] to grab the URLs and then grabbed the articles text and information via HTML parser, modified from Dr. Ryu's [1] code. All the information grabbed is stored in appropriate JSON files to be utilized in model development. This enables us to have a defined set of data, which we can preprocess and utilize in our model development of sentiment analysis and investment returns.

Figure 2 shows a sample snapshot of FNGU market data acquired from yfinance [3].



```
Bulk data tickers info:  Price                           Adj Close                                                                             ...   Volume
Ticker                          AAPL        AMZN        AVGO        CRWD        GOOG        META        MSFT        NFLX  ...       AVGO        CRWD       GOOG      META      MSFT      NFLX      NOW       NVDA
Date                                                                                                                        ...
2013-01-02 00:00:00+00:00  16.705698   12.865500    2.395675         NaN   17.969599   27.915949   22.451811   13.144286  ...   36716000         NaN  102033017  69846400  52899300  19431300  1525500  478836000
2013-01-03 00:00:00+00:00  16.494843   12.924000    2.408191         NaN   17.980036   27.686640   22.151047   13.798571  ...   23295000         NaN   93075567  63140600  48294400  27912500   863100  298888000
2013-01-04 00:00:00+00:00  16.035383   12.957500    2.392730         NaN   18.335327   28.673666   21.736479   13.711429  ...   27113000         NaN  110954331  72715400  52521100  17761100  1034400  524968000
2013-01-07 00:00:00+00:00  15.941059   13.423000    2.379478         NaN   18.255327   29.331686   21.695829   14.171429  ...   16822000         NaN   66476239  83781800  37110400  45550400  2517000  610732000
2013-01-08 00:00:00+00:00  15.983952   13.319000    2.363281         NaN   18.219299   28.972767   21.582024   13.880000  ...   20969000         NaN   67295297  45871300  44703100  24714900  2826800  466424000
...                              ...         ...         ...         ...         ...         ...         ...         ...  ...        ...         ...        ...       ...       ...       ...      ...        ...
2024-10-28 00:00:00+00:00 233.399994  188.389999  172.020004  301.320007  168.339996  578.159973  426.589996  749.119995  ...   13195800   3464800.0   20858300  10925100  14882400   2862400  1238500  173586700
2024-10-29 00:00:00+00:00 233.669998  190.830002  179.240005  310.940002  171.139999  593.280029  431.950012  759.440002  ...   23702300   3454800.0   28916100  13019100  17644100   3660400  1641100  157593600
2024-10-30 00:00:00+00:00 230.100006  192.729996  176.639999  307.450012  176.139999  591.799988  432.529999  753.739990  ...   17911500   2083800.0   49698300  26864900  29749100   1722000   950800  179418100
2024-10-31 00:00:00+00:00 225.910004  186.399994  169.770004  296.869995  172.690002  567.580017  406.350006  756.030029  ...   26086600   3475700.0   32801900  26838400  53971000   3057700  1333300  270039600
2024-11-01 00:00:00+00:00 222.910004  197.929993  168.919998  303.130005  172.649994  567.159973  410.369995  756.099976  ...   17825100   2974300.0   21626900  15267400  24209400   2995800   985300  202737100
```

Figure 2: Sample snapshot of FNGU market data

2.2 Data Pre-Processing

2.3 Data Classification and Sentiment Analysis

Model/Evaluation Metric	Accuracy	Sensitivity or Recall	Specificity	Precision	Harmonic mean
Bernoulli Naïve Bayes	86.18%	21.89%	94.88%	36.65%	27.41%
Logistic Regression	88.30%	17.63%	97.87%	52.78%	26.43%
Default MLP	88.74%	41.56%	94.55%	50.79%	45.71%
Adjusted MLP	90.05%	34.14%	97.61%	65.95%	44.99%

Table 1: Classifiers' Evaluation Metrics

2.4 Model Development for Investment

2.5 Parameter and Feature Adjustment to Maximize Investment Returns

3 Conclusion and Future Work

4 Acknowledgements

I would like to acknowledge Dr. Jin [4] for ML, classification and probability distribution concepts, Professor Avery [5] for MLP concepts,

References

[1] C. Ryu, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.

[2] scikit learn, "train_test_split," https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

[3] Y. Finance, "yfinance," https://pypi.org/project/yfinance/.

[4] R. Jin, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.

[5] K. Avery, California State University, Fullerton, https://www.fullerton.edu/ecs/cs/faculty/.