COMP4801
Detailed Project Plan

Fintech Application:
Market News Analysis and
Accelerated Stock Price Prediction

Chiang Cheng-Ru, Huang Hsiang-Jui, Wang Ching-Yuan

Supervisor: **Dr. S.M. Yiu**

Oct.1, 2018

Table of Contents

1 Project Background

With the data analytics fields advancing extremely rapidly in recent days, some concepts, such as Machine Learning, Natural Language Processing, have been heavily practiced in order to procure more precise and desirable outputs when analyzing enormous sets of data. Indeed, adopting these techniques will increase the likelihood of yielding accurate predictions; nevertheless, there are also other factors involved that one should not overlook.

To build a data analysis application that is able to work independently, there are some vital components within; for example, a module responsible for obtaining data, another module working in analyzing the data, and finally, a module that speeds up the computation of the analyzing module, if necessary. This project aims to build an application that encompasses the components described above, via implementing the existing tools, integrating the mentioned modules, and further optimizing the performance. In the following paragraphs, the discussions are separated into three main sections, which are the Natural Language Processing Section, Machine Learning Section, and the Parallel Computation Section, in order to reflect the modules that the project aims to accomplish.

1.1 Module 1 - Natural Language Processing

The utilization of Natural Language Processing technique will enable the automation of stock price prediction via textual analysis. When it comes to the prediction of the financial market, two approaches, fundamental and technical, are concerned.

The technical approach leverages the quantitative historical market data to predict the movement of future price . However, the belief of the pattern existence in market graph can be controversial and may be context-specific[1]. To illustrate, Random Walk Theorem states that stock market prediction is impossible simply based on past stock price[2]. On the other hand, fundamental analysts make assumptions by looking into different sources. These sources are composed of information regarding geopolitics, financial environment and business principles in general. Fundamental data may come from structured and numeric sources, such as macro-economic data or official financial reports. As opposed to the structured sources, the unstructured format is more available due to the vast amount of resources in the market. The most common examples are the financial news exposed to public in daily life. Fundamental analysts consume such data sources to predict the stock price.

Due to the textual analysis being industry specific, the scope of this module is to be narrowed down. With the news regarding the US energy sector being sufficient and relatively unambiguous, the stock price of the US energy sector is chosen.

1.2 Module 2 - Machine Learning

Continuing on the aforementioned discussion on the differences between fundamental and technical analysis, forecasts on the financial market have been conducted using different approaches. In the initial phase, basic statistical analysis were adopted on the time series data, such as the closing prices. As technology advances and more data become available, more variables, including the macroeconomic data like GDP, were selected. Nowadays, it is believed that the investment behavior is highly influenced by the market news and human emotions, thus rendering machine learning the most popular tool for prediction. Machine learning, by its definition, is the study of algorithms that enables the computer to learn without being explicitly programmed. Within this field, various algorithms have been adopted on the textual data. For instance, Thomas and Sycara[3] have tried to perform classification using the number of postings and the frequency of words. On the other hand, Fung et. al.[4] have conducted SVM analysis on textual news articles so as to classify the stock price between the rises and the drops. Besides, considering the complexity of financial-related data, various deep learning techniques have also been introduced. For example, Hsieh et al.[5] have applied Recurrent Neural Network on stock price prediction based on various stock indexes.

With the data coming from the Natural Language Processing module, along with the historical stock prices, predictions are to be made on the selected variables using various machine learning techniques. Based on the prediction, the effectiveness of each algorithm can be examined and the key variables can be determined in deciding the stock price of US energy sector.

1.3 Module 3 - Parallel Computation

Speaking of the module that speeds up the computation, it is where the parallel computation kicks into play. In the data analytics field, parallel computation is essential because of its nature of the mathematical operations involved - mostly matrix based computation. In such contexts, Graph Processing Unit, also known as GPU, is used as an external device that shares the computation with CPU, to achieve better performance. Since the amount of the processing cores in GPU exceeds that of CPU, GPU can achieve much better performances than CPU. As such, to develop this module, efforts will be spent on investigating ways of distributing the CPU workload to GPU in a parallel fashion, especially on ways to manage and utilize memory more efficiently. Although there are already a few well established frameworks, such as TensorFlow, that do data analytics comprehensively, given that the software optimization being hardware dependent, working on fine tuning the parameters and also the interfacing with the GPU programming layer can be done to further accelerate the computation speed.

2 Project Objectives

The project is planned to be separated into 3 modules, which are the natural language processing section, machine learning section, as well as the parallel computation section. The objectives for the respective module are listed as the follows.

- Natural Language Processing Section
    - Text-mining processing
        - Retrieve sufficient and reliable market news from online sources
        - Map the market news to the target stock
    - Utilization of semantics
        - Integrate the semantics through implementation of ontologies and dictionaries
        - Improve coreference resolution, which refers to the situation when multiple words sharing similar meanings or concepts
    - Utilization of sentiment
        - Perform the sentiment analysis and assign a reasonable weighted score to each feature group
- Machine Learning Section
    - Feature Selection
        - Select the most relevant features for further predictions
    - Utilization of historical stock prices

        - Perform regression / classification on historical stock prices and make future stock price predictions
    - Utilization of (textual) market news
        - Perform regression / classification on market news and make future stock price predictions

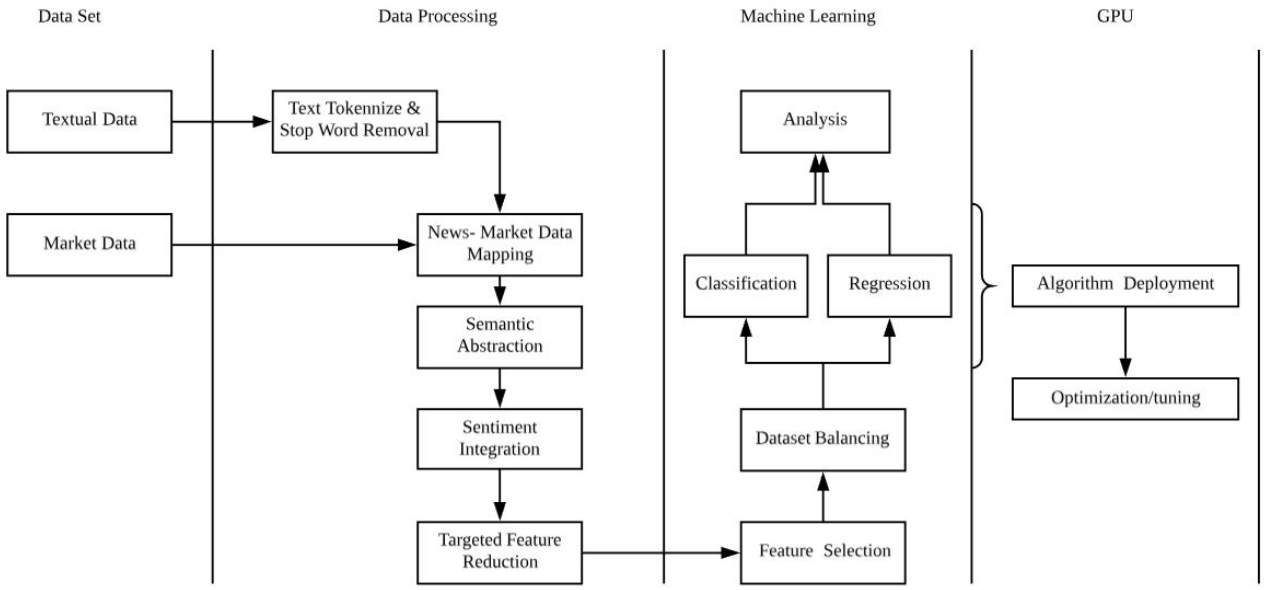
- Parallel Computation Section
    - Algorithm deployment
        - Analyze the implementation of TensorFlow, and customize it based on the algorithm adopted from the machine learning module
        - Accelerate the computation by distributing the workload to GPU
    - Speed profiling
        - Examine speed up using profiling tools

3 Project Methodology

The project comprises 3 separate modules, and each member of the team will be responsible for one of them. The work division is listed as the follows.

- Natural Language Processing - Wang Ching-Yuan
- Machine Learning - Huang Hsiang-Jui
- Parallel Computation - Chiang Cheng-Ru

The Methodologies adopted will be different for the three modules that will be developed, and they are elaborated as the follows. As with the Project Background section, the three modules are discussed separately. The diagram below describes the general workflow and the interfacing between different modules.



3.1 Module 1 - Natural Language Processing

The tools leveraged to develop this module includes Natural Language Toolkit, which deals with the human language data as well as Heuristic- Hypernym and SentiWordNet. Heuristic-Hypernym is an algorithm for disambiguating a given word, and SentiWordNet is a dictionary of sentiment values that contains a positive score or a negative score whose absolute values are between 0 and 1 for each word net entry.

The procedures are as the follows:

- Data Retrieval
  The market news is collected from reliable financial new sources, such as Wall Street Journal and Financial Times, via web crawler. The stock price is collected from Yahoo Finance.
- Pre-processing
  Natural Language Toolkit will be used to clean the raw text retrieved from the online news. The texts are to be tokenized and the stop-words will be removed. Tokenizing a text stands for breaking down a sentence into several words. For example, the sentence “Today is a good day” will be tokenized as a set of words, “Today”, “is”, “a”, “good”, “day”. Stop-words is a set of words without concrete meaning, including conjunction words, transition words. Thus, after the stop-words removal, the above example becomes, “Today”, “good”, “day”.
- Semantic Abstraction

Semantics is integrated in a manner that reduces semantic redundancy. Namely, co-reference, which is the usage of multiple words for the same concept or entity. In this layer, a method called "Bag of Words" will be conducted, to represent the new text as a group of words. Each group of words is regarded as a feature, which is a word that acts as a super-category for all subordinate words.

- Sentiment Integration
  Language sentiment will be integrated in a way that the amount of emotional-charge or sentiment-load of a word is taken into consideration in weighting a feature. After the features are determined in the feature place, it is crucial to realize that they have different levels of impact. A utilized algorithm will be performed to calculate the average weight and its contribution to the predicted movement of stock price.

3.2 Module 2 - Machine Learning

Various Machine Learning libraries for Python will be adopted for this module, including TensorFlow, SciKit-Learn, and other fundamental packages. The procedures stated below repeat until the optimal result is reached.

1. Feature Identification & Selection
   After sentiment integration is performed on the textual data, it is crucial to identify and select the key features that will decide the stock price movements.
2. Dataset balancing / Training & Test split
   To avoid sampling bias, data shuffling will be done before dividing the sample data set into training set and test set.
3. Experiments
   a. Classification: To predict the price movement (up / down) for the next period.
      Possible adopted algorithms: Multi Layer Perceptron / AdaBoost / Logistic regression / Naive Bayes Regression / Support Vector Machine / Decision Trees / Random Forest
   b. Regression: To predict the price change (%) for the next period
      Possible adopted algorithms: AdaBoost / Linear regression / Support Vector Machine / Decision Trees / Random Forest
4. Analysis
   To understand whether market news is powerful enough for making stock price prediction, comparison will be done between the prediction made merely on the past stock prices (time-series data), the one made merely on textual market news and the combined one.

3.3 Module 3 - Parallel Computation

The tools and frameworks that will be used to develop this module includes CUDA, which is the Standard Development Kit for interacting with the GPUs, and also TensorFlow framework, which is used for implementing machine learning algorithms. To facilitate the development, Numba, a Python wrapper for CUDA, might also be adopted.

In order to achieve speedup in the computation, extensive studies on the currently existing framework is necessary. In this project, the TensorFlow framework will be studied in order to understand how the machine learning algorithm is deployed onto the parallel computation interface. More specifically, more effort will be dedicated to researching on the generic layer of TensorFlow that abstracts the GPU operation.

Apart from merely having the machine learning algorithm run on the GPU devices, further optimizations can still be done. Studies have shown that the computing performance on matrix operations, measured by the amount of flops, can be largely enhanced by local memory sharing within the thread blocks[6], so that the threads in execution are not required to access the global memory for data, which costs higher overhead. Also, thread synchronization can be done more frequently in order to avoid other threads waiting on the shared resources.

An iterative methodology will be adopted to conduct the development. The procedures are as the follows:

1. Investigate on how TensorFlow framework distributes its computation load to the GPUs. Since the framework has to accommodate various kinds of GPU, it is expected that there is a generic layer that handles the interaction with the GPU interface.
2. Modify parameters for training models, and develop functions that can utilize the GPU more efficiently
3. Verify the result by profiling the function and the performance of the model

Step 2, and step 3 form a loop themselves. After the preliminary investigation on the generic layer and learning how to operate the framework, efforts will be concentrated on experimenting ways of fine tuning and better the GPU utilization.

4 Project Schedule and Milestones

Sep 30th	Deliverable of Phase 1 - Project Plan - Project Website
Oct	Investigate the usage of current existing algorithms and Studies of the analysis and design of the past research - Research on related studies of natural language processing - Study how TensorFlow interacts with the GPU layer - Get hands on experiences on TensorFlow
Nov - Dec	Preliminary Implementation - Validate data sources - Build prototypes
Jan 20th	Deliverable of Phase 2 - Preliminary implementation - Interim Report
Dec - Feb	Development
Feb - Apr	Optimization and testing
Apr 14th	Deliverable of Phase 3 - Finalized Implementation - Final report

5 Conclusion

This project consists of 3 modules, which are Natural Language Processing, Machine Learning, and Parallel Computation. By performing segmentation and sentiment analysis on the market news related to the US energy sector, the selected features will then be adopted in stock price prediction using regression and classification with various machine learning algorithms. Throughout the process, the implementation of TensorFlow will be examined so as to make improvements on the computation efficiency.

6 References

[1] Yu, H., Nartea, G. V., Gan, C., & Yao, L. J. (2013). Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets. International Review of Economics & Finance, 25, 356–371

[2] Malkiel, B.G., A Random Walk Down Wall Street. 1973, New York: W.W. Norton & Company Ltd.

[3] Thomas, J.D. and Sycara, K., 2000. Integrating genetic algorithms and text learning for financial prediction. *Data Mining with Evolutionary Algorithms*, pp.72-75.

[4] Fung, G.P.C., Yu, J.X. and Lam, W., 2002, May. News sensitive stock trend prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 481-493). Springer, Berlin, Heidelberg.

[5] Hsieh, T.J., Hsiao, H.F. and Yeh, W.C., 2011. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied soft computing*, *11*(2), pp.2510-2525.

[6] Ryoo, S., Rodrigues, C., Baghsorkhi, S., Stone, S., Kirk, D. and Hwu, W. (2008). Optimization principles and application performance evaluation of a multithreaded GPU using CUDA. *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming - PPoPP '08*.