FYP Proposal

Introduction:

Stock market is one of the most challenging problems in prediction because many factors affect the stock prices and are usually unknown in advance.

In this project, we aim to develop an automatic system for stock market prediction, which combine both News Sentiment Analysis and Technical Analysis. This will help us to create a more accurate prediction than those just using the technical data. Together with the use of Recurrent Neural Network specifically using Long Short-Term Memory (LSTM), we hope to get a clearer grasp on how different financial news will affect the stock market. We want to utilize the trained model to predict the impact and duration of these financial events towards the stock prices. The input of the model will be financial news and stock price data. The output of the model should be a prediction of the stocks in the upcoming days. This enables us to perform the best action at that moment to maximize our gain or to minimize our losses. We will also incorporate financial indicators used commonly by the financial analysts such as RSI and MACD to predict the best timing for buying, holding and selling the stocks. These strategies should help improving the model and minimize the risks due to the errors and imperfections in the model due to the unpredictable factors.

The project is divided into several parts, including:

1. Data Collection
2. Data pre-processing
3. Sentiment analysis / Summarization / Event extraction
4. Model selection and Training
5. Strategy

Data collection

There are two types of data we need to collect for this project, financial news and market data.

For financial news, we will be mainly using google news to crawl historical news for each stock since google provide convenient and unlimited searches for us to crawl news. Unlike other news API, it is totally free of charge. We can search news with a custom day range while it does not limit the time range. Hence, we can easily search historical news years ago. Therefore, it is definitely one of the best choices we have got with limited budget.

Should collect as many financial, economic and political news as possible. You can extract subsets for experimentation.

Google will block robots, so the robot should pretend to be human.

For market data, we will be using the yfinance API to collect historical market data from Yahoo finance. This API is completely free and we are able to collect stock price data in 1-minute, 2-minutes, 1-hour and 1-day intervals. For data with 1-minute, 2-minutes, and 1-hour intervals, we can only retrieve data for past few months so we may need more data for our project. In order to retrieve more market data, one of us is now collecting these data in real time so we will have one more year data before we submit our

Investigate how you could use the fine granularity data, and how long the period should be, i.e., could stock prices five years ago can be use to predict today's price?

Also, while we want to implement a real-time system to predict the next prices continuously, to test an algorithm, we typically evaluate the prediction precision using a fixed dataset. Thus, you need to keep a stable stock price dataset for testing.

final project. For data with 1-day interval, we can retrieve data for more than 20 years so it should be sufficient.

Google News API: https://pypi.org/project/GoogleNews/

YFinance API: https://pypi.org/project/yfinance/

Data preprocessing

From the price data, different technical indicators were calculated using Stockstats library.

Indicators.

Name	Formula
Simple Moving Average $SMA(t, N)$	$\frac{\sum_{k=1}^{N} pc(t-k)}{N}$
Exponential Moving Average $EMA(t, \Delta)$	$(pc(t) - EMA(t-1)) * mult + EMA(t-1)\ mult = \frac{2}{\Delta+1}, \quad \Delta = timeperiodEMA$
$MACD(t)$	$EMA(t, \Delta = 12) - EMA(t, \Delta = 16)$
Relative Strength Index $RSI(t)$	$\frac{100}{1+RS(t)}, \quad RS(t) = \frac{AvgGain(t)}{AvgLoss(t)}$
Bollinger Bands	$UpperBand(t) = 20 * SMA(t, N) + (40 * std(pc))\ MiddleBand(t) = 20 * SMA(t, N)\ LowerBand(t) = 20 * SMA(t)(40 * std(pc))$
Stochastic Oscillator $KDJ(t)$	$100 \frac{(pc(t) - MIN(pl))}{MAX(ph) - MIN(pl)}$
True Range $TR(t)$	$MAX(ph(t) - pl(t); ph(t) - pc(t-1); pc(t-1) - pl(t))$
Average True Range $ATR(t)$	$\frac{13 * ATR(t-1) + TR(t)}{14}$
Williams Indicator $WR(t)$	$\frac{MAX(ph) - pc(t)}{MAX(ph) - MIN(pl) * (-100)}$
CR indicator $CR(t)$	$100 \frac{SMA(ph(t) - MIN(m, ph(t)))}{SMA(m - MIN(m(t), pl(t)))} \quad m(t) = \frac{pl(t) + ph(t) + pc(t)}{3}$

With pc=close price, po=open price, pl=low price, ph=high price and std=standard deviation.

Sentiment analysis / Summarization / Event extraction

Sentiment analysis:

From textual data, two different sets of features were extracted using the dictionary of Loughran and McDonald (2011) (L&Mc) and AffectiveSpace. In both cases, the news is transformed into sentiment embeddings.

1. The Loughran and McDonald dictionary is specific for financial applications and contains different lists of Constraining, Litigious, Negative, Positive, Uncertainty, Superfluous, and Interesting words. It proved to be effective in many research papers in the financial forecasting field.
2. AffectiveSpace (Cambria, Fu, Bisio, & Po- ria, 2015) is a vector space model, built by means of random projection, that allows for reasoning by analogy on natural language concepts. In AffectiveSpace, each concept is mapped to a 100-dimensional vector through a dimensionality reduction of affective common-sense knowledge. This procedure allows semantic features associated with concepts to be generalized and allows concepts to be intuitively clustered according to their semantic and affective relatedness.

Alternative methods that we have reviewed and considered:

Summarization:

Pointer-generator network are chosen as the model to generate abstractive summary which will be considered as an impact factor to the stock. Pointer-generator network is a traditional encoder-decoder model with copy and coverage mechanism. Copy mechanism solve the out-of-vocabulary problem (OOV) of abstractive summarization. Coverage mechanism solve the repetitive generation problem.

Therefore, the model is supposed to give a shorter version of news which can have better semantic representation.

Event extraction

Predicting events may be difficult and could be a project by itself. The baseline is to take out keywords (maybe with a bit of phrase detection and named entities recognition) without considering specific events. Then, the problem becomes learning the correlation of the keywords to positive/negative, long-term/short-term price movements.

A Document-level Chinese Financial Event Extraction System (DCFEE) is a system that perform event extraction for financial event in Chinese word version. Event extraction are expected to be more useful in event analysis because it can extract main event that will ~~be~~ affect the stock market without any redundant information. The paper provided an application website:

http://159.226.21.226/financial_graph/online-extract.html

References:

Loughran, T. , & McDonald, B. (2011). When is a liability not a liability? Textual anal- ysis, dictionaries, and 10-ks. The Journal of Finance, 66 (1), 35–65 .

Poria, S. , Cambria, E. , Bajpai, R. , & Hussain, A. (2017). A review of affective com- puting: From unimodal analysis to multimodal fusion. Information Fusion, 37 , 98–125 .

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers, pages 1073–1083.

DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data (ACL 2018)

Model Selection and Training

Baseline Model: Technical analysis and sentiment embeddings for market trend prediction

For model selection and training, we will use Long Short-Term Memory (LSTM). This kind of Recurrent Neural Network is capable of “memorizing” and “forgetting” by its property of “input gate”, “forget gate” and “output gate”. We also use back-propagation through time (BPTT). The advantage of using LSTM is that it is easy and fast to train. This allows the model to be trained quickly while achieving acceptable results.

We will also use a technique called increasing window cross-validation for training, which should be effective for time series problems. The training phase is divided in different folds. In each step, the training data is increased by one-fold and the validation set is moved forward in time.

Reference:

Andrea Picasso, Simone Merello, Yukun Ma, Luca Oneto, Erik Cambria (2019). Technical analysis and sentiment embeddings for market trend prediction

Strategy

We aim to create a system providing a portfolio management service which can maximize the profits in the acceptable risk level of the user. There are 3 major parts in this system: user information collection and analysis, buy-in recommendation according to the acceptable risk and sell-out warning based on the real-time stock analysis.

Buy/Sell decision cannot be simply based on prediction. It depends on confidence, transaction fee, cash, etc.
See the FYP report: DL6: Memory-based Recurrent Reinforcement Learning for Systematic Trading (2017-2018)

In the user information collection and analysis, the system collects the information in the user registration process and analyses the information to generate an acceptable risk index for the recommendation. Therefore, we recommend the user updating the information once it changed so the system can provide the most suitable financial advices.

In the buy-in recommendation, the system generates a portfolio with different stocks. The portfolio will suit the acceptable risk level and investment goals of the user according the expected profit and the risk index of the stocks. Users can select the portfolio or stocks and the system will keep track on the selected stocks.

Expected Profit:

$$\text{Expected Profit} = \text{Predicted growth of the stock} + \text{Predicted dividend yield}$$

In the system, we focus on the long-term investment. In a long-term investment, the profit of buying a stock is not just from the volatility of the stock price but also the expected dividend received in the investment period.

For long-term investment, the positivity of the investment environment is very important.
To have long-term gain, the stock must be strong in itself (fundamental analysis), the overall investment environment must be positive, and the business of the company can benefit from the environment.
Thus, to predict the stock trend, we need to predict the positivity of the environment, which, I think, can be done by news analysis (e.g., covid-19, trade conflict, military conflicts, etc.)

Risk Index:

The base line of risk index is in proportion to the percentage of the expected profit from the dividend and inversely proportional to the percentage of the expected profit from the stock volatility. There are other factors affecting the risk index like the sector of the stock

In the sell-out warning, the system monitors the expected profit of every stock. Once it drops rapidly, the system an email or an app notification to users selected this stock.