THE UNIVERSITY OF MELBOURNE

COMP90019 Distributed Computing Project

Project Report

Stock prediction based on stock indicators with LSTM and stock-affecting news with BERT WWM

Tong He	**867488**
Yao Wang	**869992**

Supervisor: Richard Sinnott

Conventional Research Project (25 pts)

Declaration

We certify that

- this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university, and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

- where necessary I have received clearance for this research from the University's Ethics Committee (Approval Number ....) and have submitted all required data to the School

- the thesis is 7577 words in length (excluding text in images, table, bibliographies and appendices).

Wang Yao 王尧

Signature: He Tong 贺彤

Date: October 24th, 2019

Acknowledgment

Here we would like to express my deep and sincere appreciation to my project supervisor Professor Richard Sinnott for this kind help in my research project. Richard pointed out what the algorithms of our second model did not take into account and through this, we optimized our algorithms and the accuracy of the model has been improved. Besides, heartfelt thanks go to my parents for their verbal encouragement and financial support. Finally, we would like to thank each other that our research would have been impossible without tacit cooperation and mutual encouragement during the research.

Abstract

There are numerous studies for years struggling to make a breakthrough in stock prediction and machine learning are considered as a promising tool to take on this challenge. In this paper, we present two binary-classifier models for predicting stock price-up or price-down: the first one is based on stock indicators with LSTM and MLP for stock market indices, and the other is based on financial news with BERT WWM for a single stock. Our first model aims to predict stock fluctuation after some days with artificial neural networks based on the historical price, alongside fundamental indicators and technical indicators. For this purpose, a series of experiments were implemented to find models with the best performance for stock prediction and a backtesting system was built to simulate the real-world stock trading activities in that presenting returns the model can bring investors and producing corresponding investment strategies. The second model was built from another perspective, that is, investors' reaction to the financial news affects their trading activities thus leading to stock price fluctuation. New content pre-processing in the model involves the cutting-edge NLP model BERT WWM and also then built into an MLP-based binary classifier. In Model I, the historical stock price data of China A-share stock market were engineered into 20-days-long sequences (13337 entries in total) with 34 features and 5-day binary (price up or price down) labeling. Model II employed 44091 stock-affecting news in total for training and testing. The accuracy of both models outperform the 'buy and hold' baseline - 67.1% in Model I and 53.1% in Model II, and the backtesting system visualizes how much profit an investor can make via following trading decisions made by the modes.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Human cognition of patterns of stock market volatility is an abyss level difficulty. The essence of stock price fluctuations is supply and demand - more investors buy stocks than those who sell them leading to an increase in the stock price and conversely leading to a decrease. Traditional researchers believe the stock market is unpredictable as they put forward various theories to support this conclusion, among which Random Walk Theory and Efficient-market Theory are well-known. Efficient-market Theory (Burton G. Malkiel, 1989) states all currently observable information will reflect on the stock prices, so there is no way to predict price movement without knowing future newly revealed information. Random Walk Theory also proposed by Burton Malkiel (1999) reveals that the stock price walks a random step away from its last price, which is independently along with identically distributed in size; therefore we cannot predict price movement. The opposition believes that stock movements have some patterns that can be captured as the phenomenon that they utilize myriad methods and technologies that purportedly help gain future prices. Current methodologies for stock prediction which find favor among financial academics are indicator analysis from experts including fundamental indicator analysis and technical indicator analysis that will be detailed in this report, in the technological realm, as well as machine learning technologies and among these, artificial neural networks (ANN) are considered to be most prominent. Therefore, our research revolves around ANN from two perspectives: historical price pattern capturing and text mining; as such, two models for China stock prediction were built: one employs LSTM and MLP to process historical price data in the form of heuristic fundamental indicators as well as technical indicators. This model relies on the hypothesis that the mathematical laws of stock movement can be captured from historical prices as well as the company’s intrinsic value with ANN to predict future trends. The other model is inspired by the hypothesis that collective mood on newly released financial news will affect stock price fluctuation, thus processing financial news content with BERT WWM, an updated version of BERT - a state-of-the-art and outstanding language model developed by Google - to specify Chinese NLP tasks. In addition to these two models, a backtesting system was developed to simulate real-world trading activities of investors, thereby quantitatively exhibiting the returns these models can bring us.

This paper first will provide a basic background of stock involved in the project including the Chinese stock market, index, mains analysis paradigms. Then the cutting-edge papers on stock prediction based on ANN techniques will be discussed in Chapter 3, along with what our models draw on, differ from or optimize. Chapter 4 indicates dataset information including source, size, categories, and meanings. Later, our two models are presented separately in Chapter 5, both of which are explained in the order of the model architecture, principles, feature engineering, neural networks training, and backtesting system. In Chapter 6, the results and analysis are based on defined metrics and the simulating performance from the backtesting system. In the end, a brief conclusion and future work will be demonstrated in Chapter 7.

Chapter 2. Background

In this chapter, we will explain some of the basics of stock that need to understand before reading this paper, including the introduction of the stock market, China A-share, stock market index, two types of stock indicators their analysis paradigms and stock news with the intent to set the foundation for subsequent stock prediction content.

2.1 Stock market

‘Stock market’ refers to a public location in which shares of public listed companies are traded at an agreed price. The stock market contains three elements: marketplace, exchanger and trade variety.

'marketplace' includes institutionalized formal exchanges or over-the-counter (OTC) where a defined set of regulations are implemented (Setty, Rangaswamy, & Subramanya, 2010). Marketplace sets prices according to supply and demand where a stock that is in great demand will rise in price whereas being heavily sold will cause the price to plunge. They have been guaranteeing a secure, transparent and regulated environment in which exchangers are able to transact their assets with the lowest operational fees and risk-free. There can be many trading venues in a country or a region that may list various stocks.

'exchangers' are divided into sellers and buyers: sellers are those companies that are permitted to be traded in the marketplace, termed as 'listed company'. Listed companies first float stocks to the public. They may also offer new, additional stocks via rights issue or follow-on offers at a later stage. Initial or later issuing gives the company the ability to quickly access capital from the public. Buyers including individual retail investors and institutional investors, manage to conduct, for example, data-driven, model-driven or experience-driven analysis of the stock market to determine which stock or which type of stock portfolio to trade at that time, thus achieving profitability.

‘trade variety’ has only stocks when speaking of 'stock market', which is distinguished from 'capital market' in which other financial securities can also be traded in stock exchanges such as corporate bonds, commodity futures, etc.

2.2 China A-share

China A-shares, also known as Chinese domestic shares, are the stock shares of mainland China corporations that are denominated in RMB (Chinese Yuan) and are traded on the two Chinese stock exchanges, the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE). China A-share has the following constraints: 1. 'T+1 settlement rule'. 'T + N' abbreviation means the settlement date of trading where T denotes transaction date and N indicates the settlement of tradings is only available after N opening days (non-weekend, non-holiday) after T. Compared to 'T + 0' which is mostly used in the global stock market, 'T+1' rule prevents investors from trading same day, thus reducing large price fluctuations; 2. Daily price capping: general ±10%. The price limit is equal to the previous closing price × (1 ± price gains/losses limit percentage). Once reaching the limit in a stock, trading is halted that day; if the market is under stress, a considerable number of companies will hit this limit, inducing a serious liquidity constraint; 3. Maximum Order Size: 1 million shares.

China's market capitalization is second only to the U.S, however, A-shares is distinct from the U.S stock market. Unlike the latter which is dominated by mature institutions, the former is dominated by retail investors holding 75% of the market. Retail investors are generally irrational, trading stocks driven by emotion and speculation rather than fundamentals. Such behaviors bring chances for machine learning research on stock prediction through a certain degree of regular behavior.

2.3 Stock market index

Stock market index is a measurement of a portfolio of securities trading on a particular market. It aims to track a given market or a segment of market relying on how broad the index is (that is, which stocks are index composed of) and what algorithm the index is computed by (generally a weighted average of the market price or market capitalization of the selected stocks). The method of index construction should be clear and transparent.

Stock market index includes national index and specialized index. The national index is usually composed of the stocks of largest companies listed on a given nation's largest stock exchanges, reflecting the overall performance of the stock market of the nation, equal with investor sentiment on the economy of their country. The specialized index is compiled for describing and tracking the trend of specific sectors of the market. For example, Morgan

Stanley Biotech Index is computed from 36 American firms in the biotechnology industry, which is a benchmark used by investors to analyze the performance of biotechnology and to compare the return on a specific investment.

The purpose of our first model is to predict stock index instead of a single stock, so each stock market index is treated as a single stock which is inputted to the neural networks. The four stock indices, CSI300, CSI 500, SSE Composite and SZF200 which best reflect the performance of China's stock market were utilized to train the model:

- CSI 300 Index is one of the most commonly used 'national' indices in the China A-share stock market. It is compiled by China Securities Index Co., Ltd, composed of the 300 largest and most liquid A-share stocks traded in both SSE and SZSE to reflect the overall performance of the A-share market. According to the latest report from China Securities Index Co., Ltd, Constituents Total Market Cap of CSI 300 until September 2019 is 30424.9 billion, accounting for 54.4 fo the total market of the A-share market, among which the Constituents largest is 1491.0 billion, the smallest is 13.6 billion.
- CSI 500 Index is composed of the largest remaining 500 A-Share stocks after excluding both the CSI 300 Index constituents and the largest 300 stocks, showing the condition of small-mid caps in China.
- SZF200 Index selects 200 listed companies with the highest fundamental value as samples, which is measured by four financial indicators and determines the weight of the sample stock.
- SSE Composite Index is composed of all listed stocks on the Shanghai Stock Exchange, representing the region environment.

The compilation of these four indices can reflect the overview and operation of stock price movement in China stock market and can serve as an evaluation criterion for investment performance, providing a basis for index-based investment, for the reason that: 1.all companies with strong profitability are included and the assets of the constituents are excellent; 2. distribution of index industry is basically consistent with the sector breakdown in China; 3. the weighted average growth rate of the main business income and net profit is obviously larger than the market average.

2.4 Stock analysis paradigms

There are two accepted stock analysis paradigms one of which an investor may hold to determine their stock trading activities, that is, fundamentals indicators analysis and technical indicators analysis.

2.4.1 Fundamental indicators analysis

Fundamentals of stock refer to the intrinsic value of its issuing corporation, which includes macroeconomic together with intrinsic company context (condition of company resources). Fundamentals indicators are designed to measure the intrinsic value qualitatively and quantitively based on related data (Soni, n.d.). Fundamentals indicators include:

- Qualitative information that cannot be shown in numerical terms however the link with the nature of the corporation itself is significant as well in forecasting potential of the corporation. It may include business philosophy and model, performance and experience of the management system, talent composition, branding, competitive advantage and so forth.
- Quantitatively, the basic data (revenues, earnings, assets, debts) from annual reports of the company are typically combined to generate financial ratios as fundamental indicators - Market share, Dividend yield, Price-to-book ratio (PB), Price to earnings ratio (PE), Return on equity (ROE), etc.

Fundamental indicators analysis is based on the hypothesis that a stock price will ultimately reflect its intrinsic value over a certain time. To be more specific, it takes that any stock may wrongly be priced within a short time but will eventually fluctuate to the right price. As such, investors can profit by buying stocks at a wrongly low price and then waiting for the market to realize its mispricing and reprice the stock.

In fact, the analysis of causes of stock volatility based on fundamentals indicators is empirical, rendering it complex to combine them with machine learning algorithms for trading automation and higher prediction accuracy. Experts, for example, have formed their own experience after years of research on financial report analysis, and they may hold diverse insights on one report. Some of them can quickly find the core points in the report in the short run. They may know which are the company deliberately letting to public knowledge, which is it does not want the public to know, what are the risks that the company may be confronted with in the future and which risks are affordable by the company. Some of these ideas may be fulfilled, however, such it is hard to unify these empirical methods and thus hard to incorporate machine learning algorithms.

2.4.2 Technical indicators analysis

Technical indicators are a collection of the external manifestation of stock trading data, typically are heuristic or mathematical calculations derived from the basic historical stock data: price, volume, open interest and trade time.

There are countless heuristic technical indicators on the stock market; in addition to the basic ones, our first model used six advanced technical indicators - Volume Ratio, MACD, KDJ, DMI, Chaikin Oscillator, and Psychological Line - constructed in the feature engineering phase, which will be described in detail in 'Dataset' section of Chapter 4. Technical indicators are widely exploited by investors to evaluate historical trends and ranges to predict future price movement.

Unlike fundamental indicators analysis that focuses on the listed company, technical indicators analysis only cares about the mathematical significance of the numerical changes of a stock itself. There are three premise hypotheses in this stock analysis paradigm: (1) market behaviors contain all information; (2) price changes have certain trends or laws; (3) history will repeat itself. Since (1), all other potentials factors including fundamental indicators, policy aspects, news factors can be neglected. (2) and (3) make it possible to predict future trends based on history (Murphy, 1999).

Which one or combination of paradigms that investors will choose is based on their belief in which hypothesis. No paradigm is going to be the gospel, and our work is to consistently input these indicators with different permutation and combination to train our model to the greatest possible increase inaccuracy.

Chapter 3. Related work

Since we have researched cutting-edge papers on stock prediction, considering ANN as the most promising technological realm for stock prediction and applying them heavily in our models, this chapter will discuss related work on the use of ANN for stock prediction, including the directions, the overall performance and where our work draws on, differ from or optimize them. Generally, as shown in Figure 1, there are two main research directions on stock prediction with ANN: I. stock indicators; II. stock-related text mining.

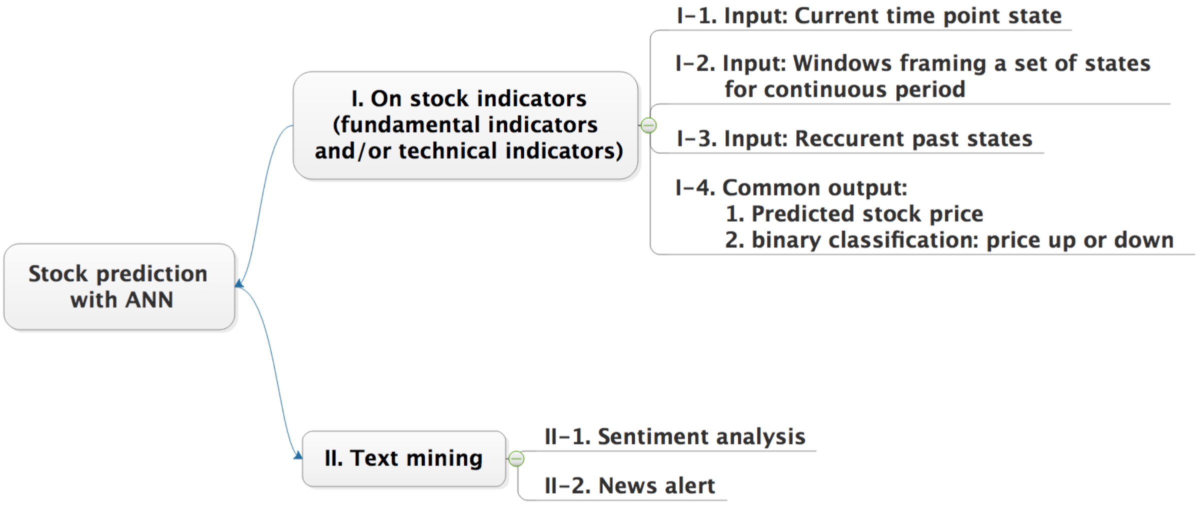

Stock prediction with ANN
I. On stock indicators (fundamental indicators and/or technical indicators)
I–1. Input: Current time point state
I–2. Input: Windows framing a set of states for continuous period
I–3. Input: Reccurent past states
I–4. Common output:
1. Predicted stock price
2. binary classification: price up or down
II. Text mining
II–1. Sentiment analysis
II–2. News alert

Figure 1 Directions of stock prediction with ANN

I is artificial intelligence for technical indicators analysis, fundamental indicators analysis or combination of them, which can be divided into three approaches according to three input patterns. I-1 aims to extract the relation between the current state point and future state point thus either to predict a specific price at a near time or to classify with a price up or down. Researchers have struggled the former for years however it concludes that until now no existing model can perform better than the one directly assigning the price in current time point to the next time point for prediction (Long, Lu, & Cui, 2019). In our first model, the fundamental indicators are entered into MLP in this approach.

I-2 means setting a window of inputs signifying a fixed set of continuous-time period states to predict the future state point; within the window, the relation between the state before and after is not passed on.

I-3 yet contains a hidden state to pass information recurrently to make ANN able to to learn the relation of states between ones from theoretically infinite time ago and future states. Nelson et. al (2017) invent an approach to stock price movement prediction using this way - they use LSTM Neural Networks for binary classification based on past price data for particular stocks in the format of basic technical indicators - time series of candles: open, close, high, low and volume - in a granularity of 15 minutes. It compares the LSTM-based model to MLP and Random Forest baselines in terms of accuracy, turning out that the average accuracy of predicting price up or down using LSTM is 55.9%, outperforming the other algorithms on average. Compared to their work, firstly our first model is based on LSTM as well, but not for a single stock, but for the index level. Secondly, the granularity of our model is 1 open-close day, not 15 minutes as China A-share follows 'T+1 settlement rule' as explained in the background section. Another reason is that fundamental indicators vectors will be concatenated with technical indicators - it is meaningless to consider fundamental indicators within a fine time granularity, i.e. short time interval, as we cannot obtain the intrinsic causes of price changes through extracting numerical variation laws with such a high frequency. Another optimizing place is that we adopt more advanced and heuristic technical indicators into our model. We are inspired by the metrics section, introducing metrics including Accuracy, Precision, Recall, F1 and average of returns to better assess our model.

The other direction - II.text mining - includes II-1. sentiment analysis and II-2. news alert. The principle of II-1 for stock prediction is that public sentiment towards stock properties or events will regularly affect the stock price, where sentiment and opinions can be mined from social media. We conducted numerous researches on sentiment analysis for stock prediction before staring building our models and the results from them are unsatisfactory (bad or tricky). The first limitation is that there exists manual labeling of whether the news is ‘good’ or ‘bad’ for stock. Papers related to II-1 collecting related text data most frequently from Twitter. Pagolu et al. (2016) collect tweets pertaining to Microsoft in 2015 - 2016, including not only ones describing Microsoft stock but also its products and services. They believe that the existing emotion classifiers are based on a different corpus, not suitable for stock prediction tasks. Thus, the first mark part of tweets manually with positive, negative or neutral for the stock as training data, and train the first classifier in which input is each tweet and output is a 3 dimensions vector. Then generate and aggregate vectors of related tweets in three days and input them into the second classifier for predicting the stock up or down on the fourth day. The paper presents 70% of accuracy for predicting stock price up or down based on the conditions that the accuracy of the first classifier is 70 % and only 355 instances aggregated from one-year tweet meaning the test instances are only about 355/10 = 35.5, which will lead to a sharp fluctuation on final accuracy by tuning. Nguyen et al. (2015) solve this difficulty by collecting the sentiment polarity from Yahoo Finance Message Board where users may annotate the message they post as one of the following sentiment tags: Strong Buy, Buy, Hold, Sell and Strong Sell. The accuracy of their original model was reduced after adding the sentiment factor as most papers do, likely for the sentiment from a small single-origin cannot represent public sentiment. Because of these limitations and unsatisfactory results, We instead undertake II-2. news alert in our second model, by directly aligning news with price up or down and corresponding active period of time. The alignment method was inspired by Gidofalvi and Elkan (2001).

Chapter 4. Dataset

The meanings of the data are mainly explained in Chapter 2, and this chapter will introduce the data sources, size, the three categories of data content we used in different models.

4.1 Data source

Collecting and ETL stock data from numerous origins is difficult and time-consuming from scratch. Stock analysis has been receiving high attention, with that corresponding big data platforms of the stock market are mature in China. Our raw data is retrieved from Tushare, a financial data platform that mainly takes on collecting, cleaning and processing financial data. It helps financial analysts focus more on the research and implementation of strategies and models. We selected the stock-related data on the Tushare, read and stored it in our local database. In Model I, the historical stock price data of China A-share stock market were engineered into 20-days-long sequences (13337 entries in total) with 34 features and 5-day binary (price up or price down) labeling. Model II employed 44091 stock-affecting news in total for training and testing.

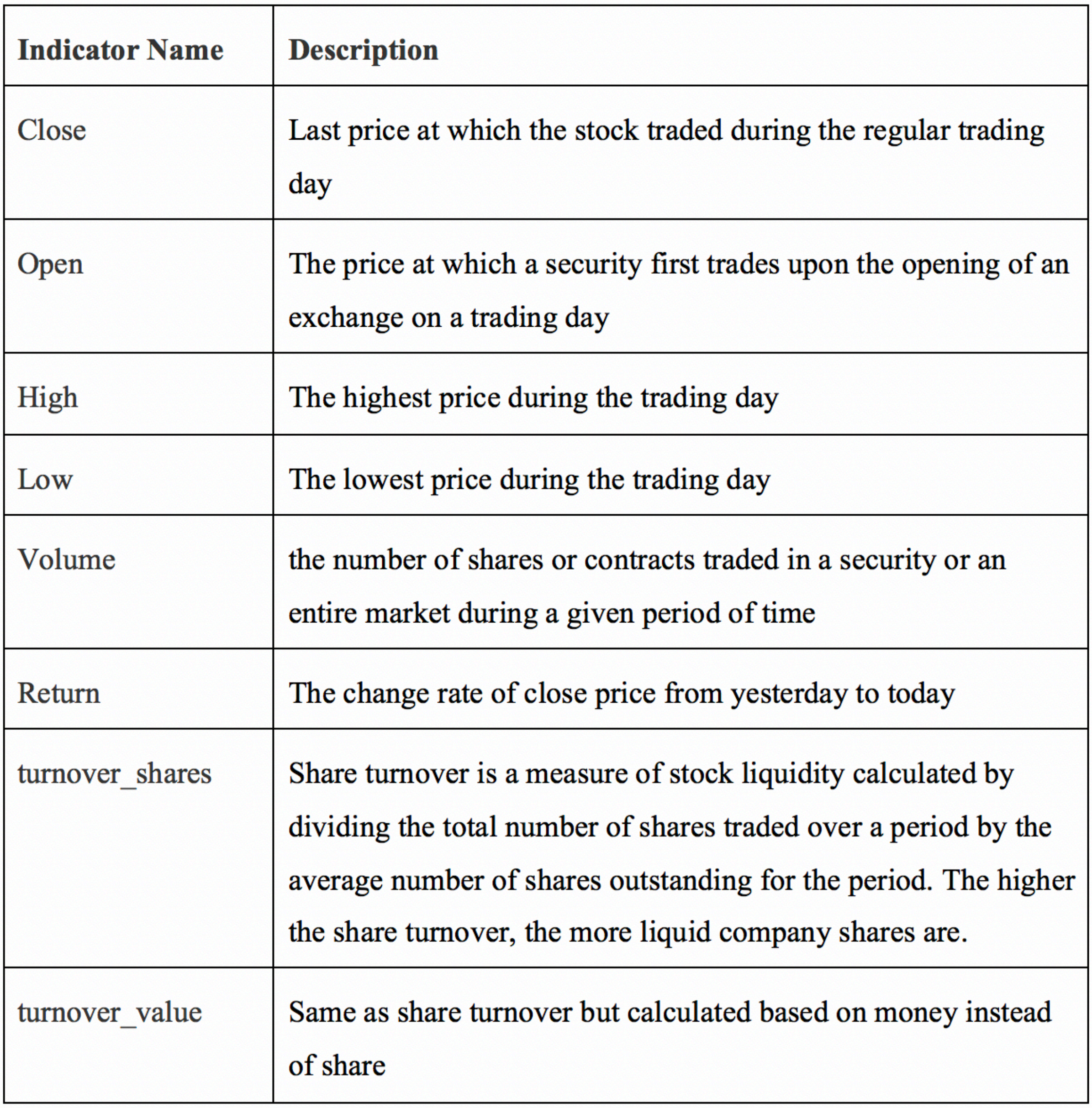
4.2 Fundamental indicators

In our first stock prediction model, we used PB and PE as two features. PB is the ratio of the stock price of a company to book value per stock. The book value of a company is its carrying value in accordance with the balance sheet, calculated as total assets minus intangible assets (patents, goodwill) and liabilities. Its pre-stock here is calculated as the total common stockholder's equity minus the preferred stock, divided by the number of common stocks of the company. PB avails investors of identifying potential investments. PE is the ratio of a company's stock price to its earnings per stock, employed by investors to assess the stock, high PE indicating the stock is over-valued or are awaited by investors high growth rates hereafter.

4.3 Technical indicators

Technical indicators focus on trading and price history, which can be expressed in statistical trends gathered from trading activities (Appel, 2005). There are 30 technical indicator features filtered in the Model I. The basic technical indicators we use are listed in Table 1 below.

Table 1 Basic technical indicators



Indicator Name	Description
Close	Last price at which the stock traded during the regular trading day
Open	The price at which a security first trades upon the opening of an exchange on a trading day
High	The highest price during the trading day
Low	The lowest price during the trading day
Volume	the number of shares or contracts traded in a security or an entire market during a given period of time
Return	The change rate of close price from yesterday to today
turnover_shares	Share turnover is a measure of stock liquidity calculated by dividing the total number of shares traded over a period by the average number of shares outstanding for the period. The higher the share turnover, the more liquid company shares are.
turnover_value	Same as share turnover but calculated based on money instead of share

4.4 Stock-affecting News

We often hear a certain company just developed a new technology and soon there comes a stock surge of its company; or the CEO of one company is reported on scandals, causing the stock price to plunge. These types of news always have an impact on the stock price movement. Specifically, news content that is sensitive to stock is shown in Table 2 below.

Table 2 News content types that are sensitive to stock

News content affecting the whole market	Relative policy changes (interest rates, monitory or fiscal policies) Government change Force majeure factors (storms, hurricanes, low rains...)
News content affecting a particular stock	The company's earnings and profits reports Changes in management Launch of new product or features Bagging of large contracts Financial scandals, court cases, patents Also the same types of news content about competitors.

We obtain news data from the Tushare's news section, which includes all of the new content types mentioned above. News within this section come from the five largest portals of Chinese stock including 10jqka, Eastmoney, Sina, WallstreetCN and yuncaijing - these five portals guarantee wide coverage, timeliness, and authenticity of the news. Each news was divided into two sections: title text and content text.

Chapter 5. Architecture and Methodology

This chapter will describe the architecture and methodologies of the two models respectively and a backtesting system for real-world trading simulation. In order to make the structure of this chapter more clearly, the content involved in each section including concepts, principles, structures, algorithms, and formulas will be explained in the order of the flowchart representing model architecture.

5.1 Model I

LSTM on stock indicators is the first model is our project, aiming to employ LSTM and MLP on historical daily stock data in the form of both fundamental indicators and technical indicators to predict stock by binary classifying price up or price down in five opening days. Figure 2 shows the architecture of Model I.

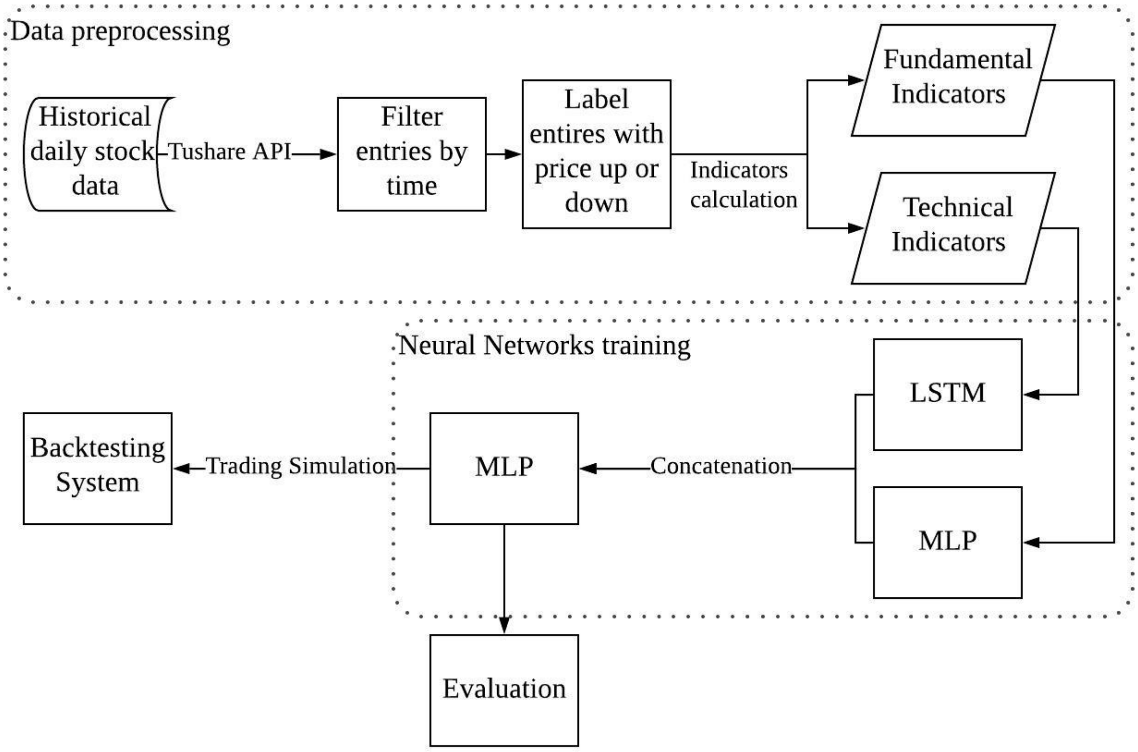


Figure 2 Architecture of Model I: LSTM on stock indicators

5.1.2 Feature engineering

(1). Data filtering

Statistically, outliers can be quite meaningful whereas our purpose is to obtain as resilient and robust a model as possible through assuming the rules of the stock market is stable, in that, the past time period we choose must be closest to what the future will be. Thus, we observed stock data from 2005 to 2019, and the changes in closing price during the period are illustrated in Figure 3. it is easy to find that the stock price has a distinct peak - a sharp rising from 2008 and a subsequently precipitous falling to the bottom in 2009. This is because of the financial crisis in 2008, leading to a large-scale decline of the stock markets - this abnormal performance cannot represent the general law of stock markets. Based on our stable stock market assumption where such force majeure factors will be considered as non-occurring in the future, we exclude this part of data from training the model to prevent the model from over- or underestimating the confidence values.

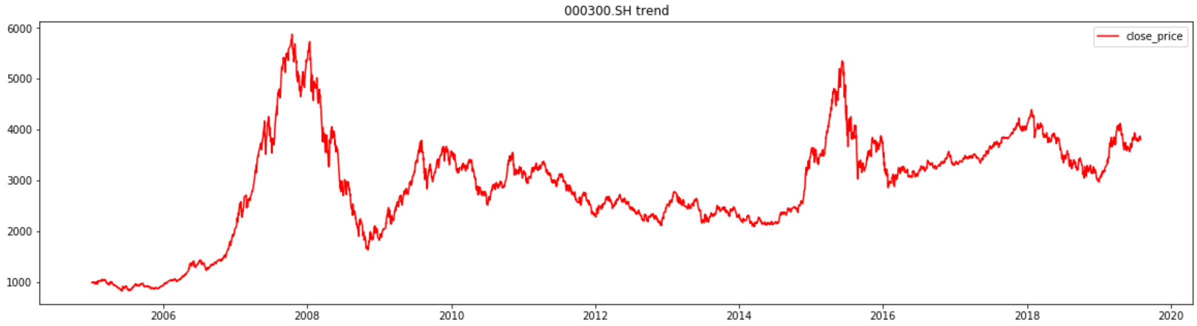


Figure 3 Close price of CSI-300 from 2006 to 2019

(2). Data Labeling

The minimum time period is set to one day as ‘T+1 settlement rule’ of China A-share mentioned before. Short-term trends have too much randomness and noise whereas the ability of the model is not enough to predict the stock market after a very long time (Pagolu et al., 2016). Observing training performance for the different intervals, we finally assigned each entry a binary class with the price trend after 5 days, '0' stands for price goes down, and '1' refers to raise up. Along the way, we use the data for the most recent year as the dev and test datasets, with the rest as the training dataset.

(3). Data Sampling

We use cross entropy as the loss function of the model witch. By observing the marked data, the proportion of price up to price down in the entire dataset is almost equal. In the training dataset, the ratio of positive to negative samples was 52: 47, thus we adjust the weights in the loss function to the reciprocal of the proportion of each category, so that the trained model conforms to the overall probability distribution as shown in Figure 4.

$$\operatorname{loss}(x,\ \text{class }) = \text{weight}[\text{class}]\left(-x[\text{class}] + \log\left(\sum_j \exp(x[j])\right)\right)$$

Figure 4 Pytorch CrossEntropyLoss function with specified weight argument

(4). Feature extraction

The raw stock data has great potential to mine. To intuitively reflect the implicit information in stock market transactions, we calculated heuristic indicators on basic data as features. These indicators algorithms are selected from the indicators we have chosen to appear in the relevant literature and in the investor forum. The table below shows our processed indicators.

Table 3 List of calculated technical indicators as new features

Indicator Name	**Explanation & Algorithm**
MACD	an acronym for **Moving average convergence/divergence**, used to identify moving averages that are indicating a new trend (Appel, 2005). $EMA_n = \text{Closing Price }_n \frac{2}{\text{Time Period }+1} + EMA_{n-1}\left(1-\frac{2}{\text{Time Period }+1}\right)$ $\text{signal\_n } = MACD_n \frac{2}{\text{Time Period }+1} + \text{ signal\_n-1}\left(1-\frac{2}{\text{Time Period }+1}\right)$
KDJ	short for **Stochastic oscillator -** where stochastic refers to the current price point along with its price range over a period of time, calculated as: $\%K = 100 * (\text{Price} - L5)/(H5 - L5)$ $\%D = ((K1 + K2 + K3)/3)$ Where H5 and L5 denote the highest and lowest prices of the 5 previous trading session respectively (Murphy, 1999).

	%D is the 3-day moving average of %K (the last 3 values of %K).
DMI	means **Directional movement index,** indicates price moving direction proposed by J. Welles Wilder (1978). $+\mathrm{DI}=\left(\frac{\text{Smoothed }+\mathrm{DM}}{\mathrm{ATR}}\right)\times 100$ $-\mathrm{DI}=\left(\frac{\text{Smoothed }-\mathrm{DM}}{\mathrm{ATR}}\right)\times 100$ $\mathrm{DX}=\left(\frac{\lvert +\mathrm{DI}--\mathrm{DI}\rvert}{\lvert +\mathrm{DI}+-\mathrm{DI}\rvert}\right)\times 100$ Where $+\mathrm{DM}(\text{ Directional Movement })=\text{ Current High }-\mathrm{PH}$ $\mathrm{PH}=\text{ Previous high}$ $-\mathrm{DM}=\text{ Previous Low - Current Low}$ $\text{Smoothed }+/-\mathrm{DM}=\sum_{t=1}^{14}\mathrm{DM}-\left(\frac{\sum_{t=1}^{14}\mathrm{DM}}{14}\right)+\mathrm{CDM}$ $\mathrm{CDM}=\text{ Current DM}$ $\mathrm{ATR}=\text{ Average True Range}$
Chaikin oscillator	The oscillator measures the accumulation-distribution line of moving average convergence-divergence (Kaufman & Chaikin, 1991). $\mathrm{N}=\frac{(\text{ Close }-\text{Low})-(\text{ High }-\text{ Close })}{\text{High }-\text{Low}}$ M =N* Volume (Period) $\mathrm{ADL}=\mathrm{M}(\text{ Period }-1)+\mathrm{M}(\text{ Period })$ $\mathrm{CO}=(3-\text{ day EMA of ADL})-(10-\text{ day EMA of ADL })$ Where $\mathrm{N}=\text{ Money flow multiplier}$ $\mathrm{M}=\text{ Money flow volume}$ $\mathrm{ADL}=\text{ Accumulation distribution line}$ $\mathrm{CO}=\text{ Chaikin oscillator}$
Psychological line	refers to the number of days that have risen within n days of the selected period (Murphy, 1999). $PsychologicalLine=100\times DUP_n/n$ Where $\mathrm{N}=\text{ time period }=10\text{ days}$ $\mathrm{DUP}_{\mathrm{n}}=\text{ Days of price increase in n days}$
Volume ratio	the ratio of the sum of daily trading volume during the period to the total volume in the period (Lukac, Brorsen, & Irwin, 1988). $VolumeRatio=100\times\sum VUP_n/\sum V_n$

5.1.3 Neural networks training

The neural network architecture in Model I is shown in Figure 5: entries with technical indicators are inputted into the LSTM neural network and the others with fundamental indicators are entered into MLP for training; then concatenate the outputs from these two networks and input them into another coupled MLP to binary classify price up or down.

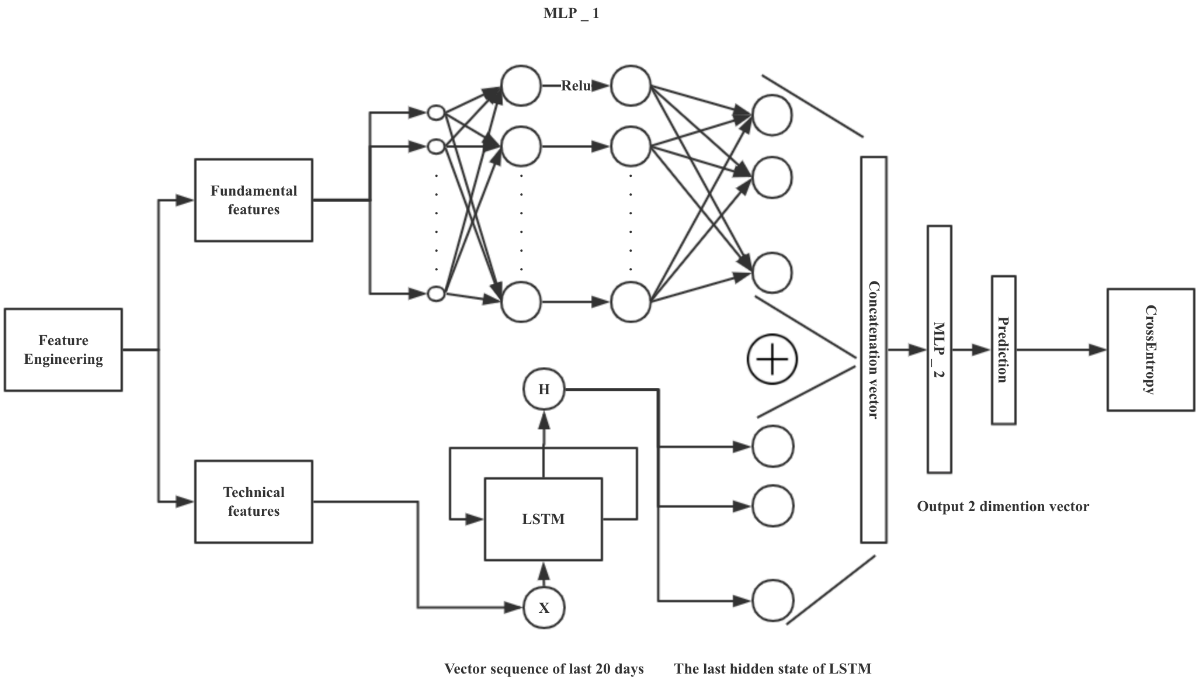


Figure 5 Architecture of the neural networks in Model I

(1). Multi-layer perceptron (MLP)

A multilayer perceptron (MLP) is a class of feedforward ANN consisting at least three (an input and an output layer with one or more hidden layers) nonlinearly-activating nodes all of which are fully connected (Each node in every layer (except for the output layer) is connected to every node in the next layer of the network) (Hastie, Tibshirani, Friedman, & Franklin, 2005). MLP learns the pattern in the dataset by changing connection weights through error backpropagation. Since fundamental indicators have slow updates and do not involve time series problems, we use the first MLP to fit the relationship between fundamental features and stock value. After the concatenation, we use the second MLP to transform concatenating vectors into two-dimensional vectors representing up and down.

(2). LSTM neural network

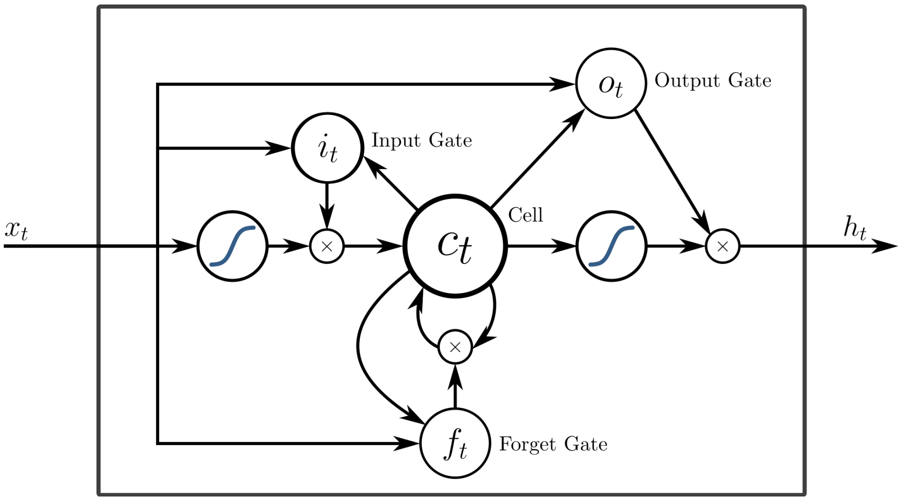


Figure 6 Architecture of LSTM

Long short-term memory (LSTM) is an ANN that can process entire data sequence with feedback connections, alongside carrying out the exploding and vanishing gradient problems happening in training traditional RNNs (Hochreiter & Schmidhuber, 1997). Figure 6 presents the graphical representation of an LSTM unit with peephole connections, where Input Gate, Output Gate and Forget Gate are leveraged to record useful state and forget unuseful state. The equations of LSTM include:

$$f_t = \sigma_g(W_f x_t + U_f c_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i x_t + U_i c_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o x_t + U_o c_{t-1} + b_o)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + b_c)$$
$$h_t = \sigma_h(o_t \circ c_t)$$

Where $i_t o_t$and f_t denote the activations of respectively the input, output and forget gates at time step t and c_t denote activation of the memory cell c at time step t-1. LSTM fits technical indicator analysis well as the key of the latter is to find the trend over time as mentioned above. The output of LSTM in our model is the last hidden state that contains the stock information within a time period.

(3). Concatenation

The outputs of the two networks are combined into one vector. It is then entered into another MLP network to produce the final result.

(4). Model training

The loss function has been shown in Figure 4 in Data sampling section. The model leverages Pytorch Adam optimizer with the learning rate set to 1e-3.

In the early stage of the study, Overfitting happened in the training set shown in Figure 7 below. One possible reason is that the amount of data divided by day is less than the amount of data divided by minutes, another is that we have introduced more indicators into the model.

To eliminate overfitting, we first introduced more indicators for training, using dropout and L2 regularization techniques on the model to achieve some effect. Moreover, we reduced the number of hidden layer nodes in the MLP_2 from 32 to 16. The simplified model loses some of its performance on the training set but has better generalization capabilities.

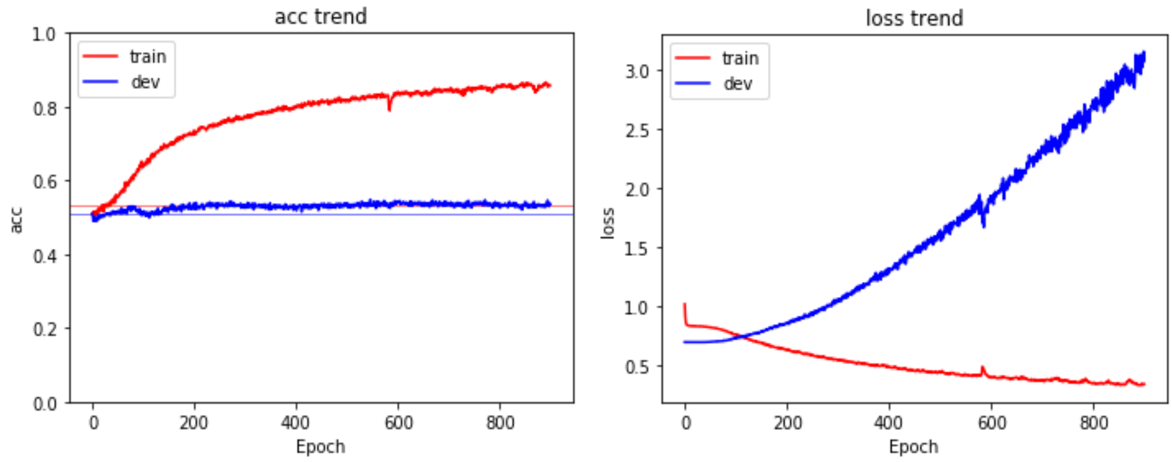


Figure 7 Overfitting on the training dataset

5.1.4 Backtesting system

The indicators mentioned above can be used to guide and monitor the training of models or to compare the performance of different models, but the ultimate goal of all quantitative transactions is to obtain more benefits. ‘Buy and Hold’ baseline was chosen for comparison, which means buy at the first time step and sell at the latest. Our backtesting system is based on it to simulate the performance of the model in actual trading. Distinguished from other quantitive algorithms such as ranking stocks, this model directly predicts whether the stock price rises or falls after 5 days in the future. Based on this we create the following strategy. When the prediction is raised up, buy or continue holding stocks for 5 days. When the prediction is going down to sell all stocks in the future and wait for the next upside opportunity.

5.2 Model II

Our second model is stock prediction ANN model based on stock-affecting news with the state-of-art language model Bert WWM. the architecture of Model II

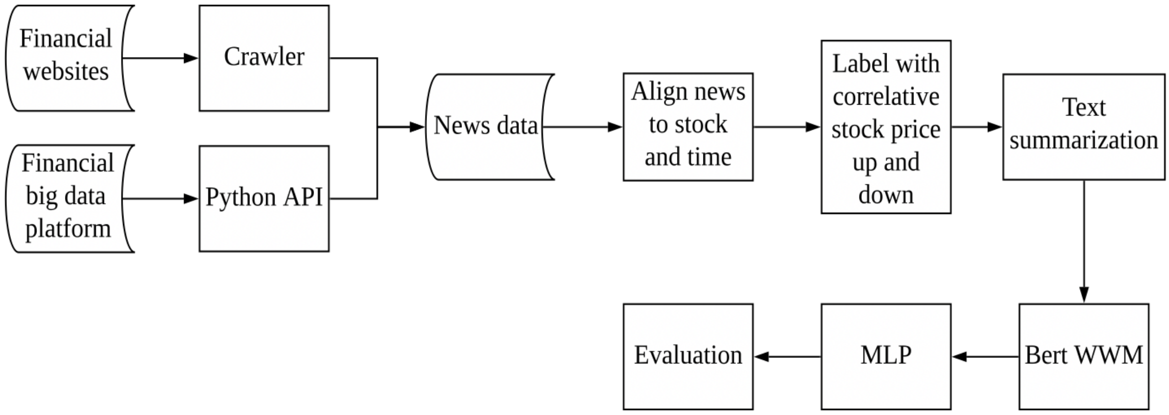

Financial websites
Crawler
Financial big data platform
Python API
News data
Align news to stock and time
Label with correlative stock price up and down
Text summarization
Bert WWM
MLP
Evaluation

Figure 8 Architecture of Model II:stock prediction ANN model based on stock-affecting news with the state-of-art language model Bert WWM

5.2.1 Principles for stock prediction based on financial news

This model relies on the hypothesis that an investor’s trading behaviors are typically affected by observable and valuable news. Negative news likely leads to retail investors to sell stocks, such as news about company financial hardship, big layoff, political uncertainty, etc., which will translate to selling pressure and dropping in stock price. Conversely, good news the company or external environment will translate into buying action and thus a price rising. However, we cannot just simply conclude that negative/positive news makes the price down/up. Investors receive news at a different time from varied news origin (official, leaked, or rumored) and then make individual reactions based on their subjective interpretations of the news. The same news may have very different reactions, depending mainly on three aspects from investors: (1) perceptions. Paradigms of Investors on how news affects the stock vary; (2) expectations. News shows, for example, that a company has better-than-expected profits, which in theory, will make the stock price increase; however if these profits were expected by most investors, the price will likely stay the same (Glaser & Weber, 2005); (3) prevailing sentiment levels. these levels involve rationality v.s. sensibility, patience v.s.impatience, good or bad mental bearing ability, etc.

Besides, investors have different contributions to the stock price, which depends on the number of stocks they hold and the trading range. Thus we cannot just simply make sentiment polarity direct correlated with the stock price. Another factor is that investors have different contributions to the stock price, which depends on the number of stocks they hold and the trading range. Thus we cannot just simply make sentiment polarity direct correlated with the stock price, instead mining patterns from the source.

5.2.1 Bert WWM

Bidirectional Encoder Representations from Transformers (BERT), a state-of-art technique for NLP pre-training, open sourced by researchers at Google AI Language in 2018, has proven to outperform/lead across various NLP tasks. It makes it true to utilize the enormous amount of unannotated web text trained on power Google GPUs called pre-training, breaking the human-labeled training limitations. The model learns by two mechanisms illustrated in Figure 9 below: (1).masking 15% words in a sentence in the input and then condition each word bidirectionally to predict the masked words; (2). making the next sentence as the label of the sentence from the web to train (Devlin, Chang, Lee, & Toutanova, 2018).

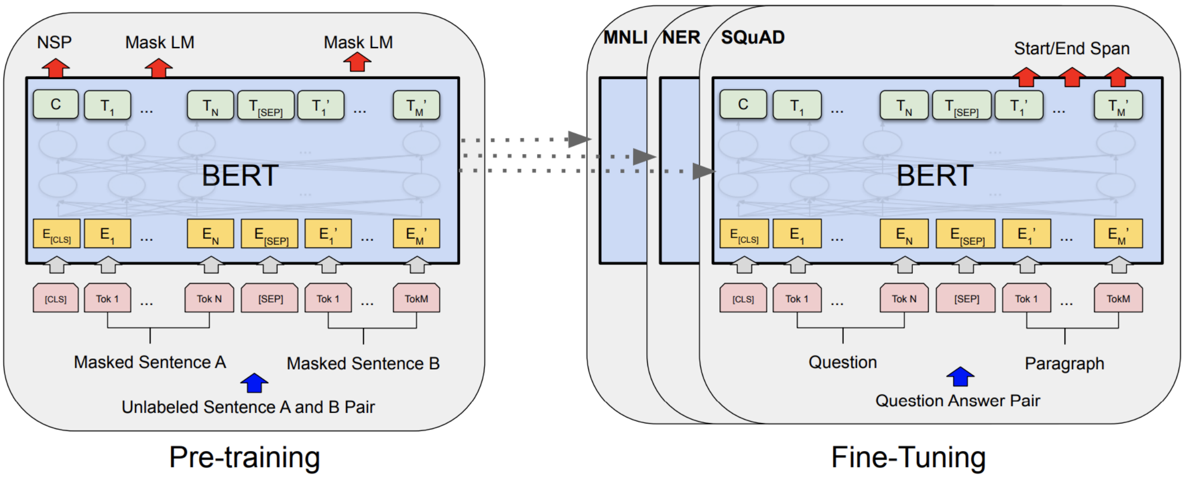


Figure 9 Overall pre-training and fine-tuning procedures for BERT.

In our model, we use Bert Whole Word Masking (WWM), an updated version of Bert for processing Chinese, released this year by Cui et al. (2019). Unlike one word is equal to one token(unit) in English, in Chinese one word has 1,2,3 or 4 tokens (units). Bert WWM implements whole word masking in Chinese text, that masking the whole word instead of masking one token. This modification has proven to outperform in Chinese NLP tasks than Bert.

In our model, we use the pre-trained model of Bert WWM to fine-tune on news text to generate embeddings and then input them into MLP.

5.2.2 Feature engineering

(1). News alignment

Labeling news with ‘good’ or ‘bad’ for the stock price up or down is crucial. Some related work label news manually, leading to a huge workload; or search sentiment tags on financial discussion board, the results of which may be incomplete and bias. In our work, we solve it by directly matching news with certain time intervals and price, called ‘alignment’ (Gidofalvi & Elkan, 2001). To be more specific, if a piece of news is released before today’s closing time, this news will be regarded affecting the next opening day’s stock price of the affected companies’stocks, and for this, if the next opening day’s closing price is higher than today’s closing price, then news will be marked ‘price up’, otherwise marked ‘price down’. If the news is released after today’s closing time, the news will be treated as released in the next opening day, thus the processing way is the same as if it was the next opening day’s news.

(2). Text summarization

Due to the limitation of hardware resources as well as 512 units of Bert maximum text length, also to reduce semantic loss from brute-forcibly truncating sentences, we use Textrank algorithm to choose the most core sentences in each news text for shortening each one into 128 units. Below is the iterative formula of Textrank:

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

where the left side of the equation represents the weight of a sentence (WS is an abbreviation for weight_sum), and the summation on the right side indicates how much each adjacent sentence contributes to the sentence. The denominator $w_{ji}$of the summation represents the degree of similarity between the two sentences. The denominator is again a weight_sum, and WS(V_j) represents the weight of the last iteration j (Mihalcea & Tarau, 2004).

5.2.3 Model training

Input preprocessed news text into Bert WWM for fine-tuning, and then the output vector which matches the class label representing the whole text is inputted into an MLP binary classifier (the structure of it is the same as the MLP in Model I).

Chapter 6. Evaluation

6.1 Metrics

In order to evaluate the performance of the model, the confusion matrix is used to visualize the result of the model output. It counts the number of correct and incorrect predictions for each class. For the 2 classification problem, it consists of 4 data, the predictions correctly made for positive classes (true positives - tp), negative classes (true negatives - tn) and those made inaccurately for both classes (false positive - fp, false-negative - fn) (Powers, 2011). We use the positive class to indicate the rise, and the negative class to indicate the fall. Thus the predictions correctly made for positive classes. Based on this we calculated 4 indicators which are accuracy precision, recall, and F1-score. Their formulas are shown below:

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$
$$Precision = \frac{tp}{tp + fp}$$
$$Recall = \frac{tp}{tp + fn}$$
$$F1\_score = 2\frac{P * R}{P + R}$$

6.2 Results and analysis

This section demonstrates the results of Model I and Model II based on the metrics listed before and the performance of the backtesting system in Model I, along with respective analysis.

6.2.1 Model I performance

In the dev dataset, the ratio of stocks' ups and downs was 47:53. To explore the impact of technical indicators, table 4 shows the Model performance using different data.

The comparison results of model are listed in Table 4, in which 'All positive' is the baseline where all the entries are given the positive label 'price up'; 'HMM' model is implemented as the binary classifier baseline for stock prediction; the feature ranges of 'Model_basic', 'Model_tech', Model_T+F' in the model we built are respectively 'only the basic technical indicators', 'the basic and the advanced technical indicators', 'all the technical indicators and fundamental indicators', in this way, we can improve the model performance by tuning features.

Table 4 Performance of different models

	Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall
All positive	47%	47%	100%	-	0%
Model_basic	53.4%	50.4%	52.7%	56.3%	54.0%
HMM	56.6%	53.6%	56.3%	59.4%	56.8%
Model_tech	58.9%	56.5%	54.1%	60.8%	63.1%
Model_T+F	**63.5%**	**61.9%**	**59.7%**	**64.8%**	**66.9%**

The results prove that adding more technical indicators actually improves the performance of the model. Compared with traditional methods, neural networks have more potential in stock price prediction, but the performance of neural networks is positively correlated with the quality of the dataset.

Table 5 shows the performance of our final Model I, that is Model_T+F for each class. For the positive class, a higher recall means capturing more profit opportunities, and a higher precision means more average return per investment. F1score is the comprehensive performance of these two metrics and 60.8% proved the profitability of MODEL_T+F.

Table 5 Performance of MODEL I

Class	Precision	Recall	F1_score
Positive	61.9%	59.7%	60.8%
Negative	64.8%	66.9%	65.8%

The figure below shows the results of the backtesting of the first model within half a year. For ease of understanding, the graph shows the percentage instead of the real price, and the value of the starting time point is assumed to be 100%.

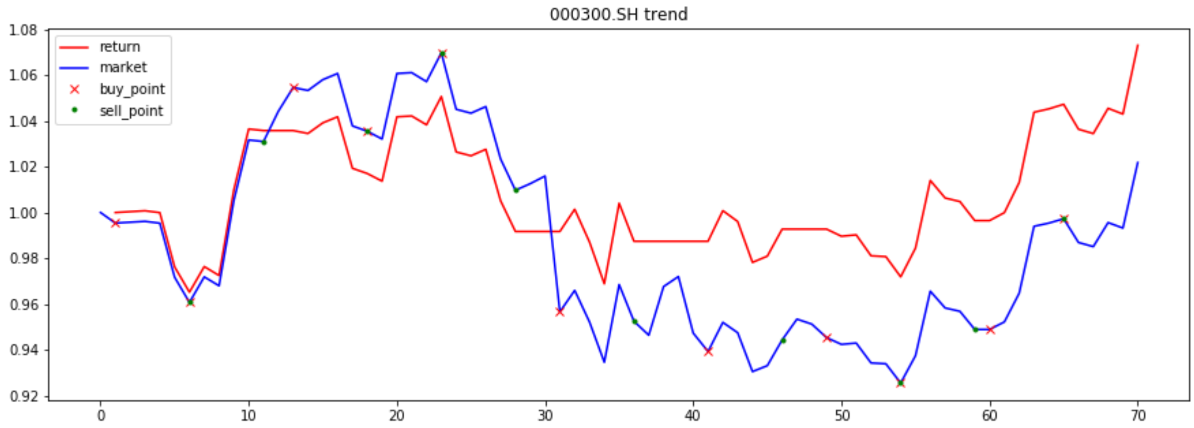


Figure 10 Backtesting of MODEL I

The blue line shows the trend of the closing price of the CSI 300, and also represents the revenue of the 'buy and hold' strategy. The red line shows the trend of investment over time. After half a year, the baseline strategy achieved a 2.1% gain, while the model achieved a 7.3% gain. The gap between the two lines refers to the relative return of model I, which is 5.2%. Because the strategy only considers one stock index, and each purchase held stock for at least 5 days, there were some fluctuations in earnings. The maximum retracement is 6%.

Because the backtesting ignores problems such as handling fees and transaction time selection, it cannot fully reflect the real situation, but the result still shows the model's profitability.

6.2.2 Model II performance

The accuracy of news prediction is 53.1%. The prediction for rising is 51.6%, while the prediction for the decline is 55.8%. This model has better judgment on negative news. The possible reason is that the impact of good news such as policies or product launches on stocks is difficult to judge directly, while negative news such as the company’s financial problems and scandals often directly affect the stock price.

Table 6 Performance of MODEL II

MODEL II	Overall	Positive	Negative
Accuracy & Precision	53.1%	51.6%	55.8%.

Chapter 7. Conclusion and future work

Overall, we have comprehensively and deeply researched state-of-art papers for stock prediction and developed two models aiming to binary classifying stock up or down in the future and one backtesting system for simulating real-world stock trading activities. Among them, Model I is designed to predict the trend of specified Chinese index, thus recommending a combination of stock in the index. Model II focuses on the single stock trend, providing support for the specified stock trending, The results prove that the stock can be predicted using ANNs to find stock patterns from historical price data or stock-affecting news, in which the model based on the former outperforms that of the latter.

We have tried to combine the two models, that is, to put indicators features and news features into the same neural network for training or to weight the results of the two models to obtain the final result, but the performance is worse than those in separate. The results of Competition 2Sigma at Kaggle also show the overall accuracy will decrease if the news feature is added to their model (Jiao & Jakubowicz, 2017) (Oncharoen & Vateekul, 2018). A possible reason is that news feature is sparse compared to stock indicators which exist in each time point. It is also likely because we cannot accurately determine after what time the effect of news will reflect on the stock price. From the perspective of investors, they are more willing to pay for two stock prediction software based on two different hypotheses rather than a mix of them as they want more suggests from diverse dimensions. Therefore, our future work is to improve our model performance and develop them as two real-time stock prediction products: one is the software for a stock recommendation based on historical price and the other is news alert software for stock monitoring.

References

Appel, G. (2005). *Technical analysis: Power tools for active investors*. Upper Saddle River, NJ: Financial Times/Prentice Hall.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., & Hu, G. (2019). Pre-Training with Whole Word Masking for Chinese BERT. *ArXiv Preprint ArXiv:1906.08101*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, *34*, 100190. https://doi.org/10.1016/j.cosrev.2019.08.001

Gidofalvi, G., & Elkan, C. (2001). Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*.

Glaser, M., & Weber, M. (2005). September 11 and stock return expectations of individual investors. *Review of Finance*, *9*(2), 243–279.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

Jiao, Y., & Jakubowicz, J. (2017). Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks. *2017 IEEE International Conference on Big Data (Big Data)*, 4705–4713. IEEE.

Kaufman, S. K., & Chaikin, M. (1991). The Use of Price-Volume Crossover Patterns in Technical Analysis. *MTA Journal*, *37*, 35–41.

Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, *164*, 163–173.

Lukac, L. P., Brorsen, B. W., & Irwin, S. H. (1988). A test of futures market disequilibrium using twelve different technical trading systems. *Applied Economics*, *20*(5), 623–639. https://doi.org/10.1080/00036848800000113

Malkiel, Burton G. (1989). Efficient Market Hypothesis. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Finance* (pp. 127–134). https://doi.org/10.1007/978-1-349-20213-3_13

Malkiel, Burton Gordon. (1999). *A random walk down Wall Street: Including a life-cycle guide to personal investing*. WW Norton & Company.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.

Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin.

Nelson, D. M. Q., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. *2017 International Joint Conference on Neural Networks (IJCNN)*, 1419–1426. https://doi.org/10.1109/IJCNN.2017.7966019

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Oncharoen, P., & Vateekul, P. (2018). Deep learning for stock market prediction using event embedding and technical indicators. *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 19–24. IEEE.

Pagolu, V. S., Challa, K. N. R., Panda, G., & Majhi, B. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *ArXiv:1610.09225 [Cs]*. Retrieved from http://arxiv.org/abs/1610.09225

Powers, D. M. (2011). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.*

Setty, D. V., Rangaswamy, T. M., & Subramanya, K. N. (2010). A review on Data Mining

Applications to the Performance of Stock Marketing. *International Journal of Computer Applications*, *1*(3), 33–43. https://doi.org/10.5120/88-187

Soni, S. (n.d.). Applications of ANNs in Stock Market Prediction: A Survey. *Engineering Technology*, *2*(3), 13.

Appendix

Code link

https://github.com/Fish-WY/stock_prediction

Work breakdown

Team member	Work
Yao Wang	Build model I Model I coding Historical price data pre-processing Develop the backtesting system Research state-of-art related papers Arrange hardware resources
Tong He	Build model II Model II coding News data ETL Thesis polishing Research state-of-art related papers