Tampere University

Heng Gui

STOCK PREDICTION BASED ON SOCIAL MEDIA DATA VIA SENTIMENT ANALYSIS

A study on Reddit

Faculty of Information Technology and Communication Sciences
Master's Thesis
November 2019

ABSTRACT

Heng Gui: Stock Prediction Based on Social Media Data via Sentiment Analysis, a Study on Reddit

Master's Thesis

Tampere University

Computational Big Data Analytics

November 2019

With the development of internet and information technology, online text data has become available and accessible for research in many fields including stock prediction. Social media, being one of the biggest content generators on the internet, is a great data resource for text mining and stock prediction. It has a large capacity, high data density, and fast information spread.

In this thesis, analyses on the relationship between the stock-related text in social media (Reddit) and the price changes of corresponding stocks are implemented. In the analysis, sentiment analysis is first applied to extract the individual users' emotions and opinions about the stocks. After that, the extracted features are analyzed via descriptive statistics and predictive analysis using the Pearson correlation coefficient and machine learning models. The predictive analysis is designed to examine the dependence between the social media text data and stock price change by evaluating the performance of predictions, four indicators are used in the evaluation including "prediction accuracy on price change direction" and three indicators in simulated algorithm trading experiments based on prediction results. They are "total profit with trading strategy for single stock", "daily profit efficiency of trading strategy" and "total profit with Portfolio trading strategy". From the results and the comparison with a Buy and Hold (B&H) baseline strategy, the predictions show good results in terms of "daily profit efficiency" and "total profit with Portfolio trading strategy". Therefore, the online forum text from Reddit are proved to be correlated with future stock price changes and might be used to make more profit than B&H strategy by incorporating their information in portfolio trading strategies.

Keywords: Stock prediction, sentiment analysis, predictive statistics, machine learning, social media

PREFACE

After overcoming all the obstacles and struggles, I am finally sitting here and writing this preface, which means my thesis is finally finished.

While enjoying the happiness, accomplishment, and release that come from completing the thesis, I would like to express the appreciation to all the people that helped me this year.

I would like to thank my supervisor Prof. Jaakko Peltonen for his professional advice, comments, and instructions.

I would like to thank all my friends, my teachers, and my family for their supporting and understanding about my abroad career.

Tampere, 13 / November / 2019

Heng Gui

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Predicting stock price is always an attractive topic for researchers. Many mathematicians, statisticians, and economists have put their effort into attempting to prove the market predictability or to forecast the prices of stock markets using different techniques and data resources. A representative milestone work comes from Fama (Fama, 1965; Fama, 1970) who suggested that the markets are unpredictable and came up with the Efficient Market Hypothesis (EMH). EMH illustrated that in efficiency markets, all market indicators including stock price would immediately adjust to any related information. Which blocked the potential prediction. Later, many researchers developed methods that have been able to successfully predict stock price, which demonstrated that the EMH was doubtable. In 2005, Lo introduced Adaptive Market Hypothesis (AMH) (Lo, 2005) for stock prediction. AMH provided a systematic theoretical explanation that the stock markets are predictable to some degree. After that, the conflict between EMH and AMH has encouraged even more research on stock prediction. Scholars have applied different techniques and theories, used different data in their research and got different results. Nowadays, the applications of stock prediction are mature. Represented by Algorithmic Trading and Quantitative Trading (Chan, 2013; Evans, Pappas, & Xhafa, 2013; Guo, Lai, Shek, & Wong, 2016; Li, Wang, et al., 2014), the new techniques have been applied in real stock markets to make profits. Their success has proven that stock predictions have the ability to make a profit. Nonetheless, research keeps working on new prediction methods and data sources in the field of stock prediction.

Since the internet was invented in 1991 (Dijck, 2013), the growth of the internet has never ended. In 2019, the population of internet users reached 7.676 billion and about 45% of them were social media active users. In the recent 5 years, active social media users hit 3.484 billion with a consistent increase from 1.857 billion in 2014 (Hootsuite

Media, 2019). People spent about 2 hours each day on social media on average. With the large population and countless content, online text data has become an important data source for researchers and practitioners. However, whether the online text is valuable for stock prediction is still under research. Some previous works (Bollen, Mao, & Zeng, 2011; Nguyen, Shirai, & Velcin, 2015; Urquhart & Hudson, 2013; Yu, Duan, & Cao, 2013) have attempted to prove that online news and social media text can influence, or can be used to predict stock price changes. Due to the potential of social media content and text mining techniques, this field still merits more attention and effort.

Most of the social-media-related works in the field of stock prediction have used text data from Twitter. However, besides Twitter, Reddit (reddit.com) is another good source of social media text data and there has been little research on using Reddit data in stock prediction. Thus, this thesis will use social media data from Reddit in stock prediction.

Reddit is an online forum that works in some ways like other social media so that users have free access to view and create posts, replies, votes, and comments in Reddit. What makes Reddit different from other social media is that the discussion in Reddit is classified into sub-boards (called ‘subreddit’) focused on specific topics.

In this thesis, experiments on the relationship between stock-related text in Reddit and stock prices are implemented. First, sentiment analysis is conducted that transformed the unstructured text data into numerical features. Then, descriptive statistics are computed in order to investigate the linear correlation between sentiment features and future price changes. After that, predictive analysis is done via machine learning methods to investigate hidden patterns. Further, an investment simulation experiment is also applied to examine the performance of machine learning models as well as the predictability of different stocks. This thesis aims to investigate whether the online forum texts in Reddit are associated with stock market price changes and whether they are valuable for stock prediction. Moreover, a profitable trading strategy is designed to prove the predictability and exploit the correlations.

The thesis is organized as follows. In Chapter 2, a literature review about stock prediction in previous research, including methods and data sources, is presented. In addition, the theoretical basis of stock price predictability is introduced. In Chapter 3, the techniques of sentiment analysis and machine learning are introduced. Chapter 4 explains why Reddit data are used and how it was collected. Chapters 5 and 6 describe the data processing and the results of data analyses. More specifically, Chapter 5 describes the feature extraction and machine learning model construction. Section 6.1 presents the results of descriptive statistics, and Section 6.2 shows the results of predictive analysis. In Section 6.3 investment simulations and trading strategies are implemented to excavate hidden correlations and exploit their prediction power for making a profit. Chapter 7 provides conclusions, discusses the limitations of the work and describes the potential directions of future work.

2 LITERATURE REVIEW

In this chapter, a review of previous research on stock prediction and sentiment analysis is presented.

2.1 Stock prediction methods

The stock market prediction has always been an attractive research field for Statistician and Computer scientists. Many scholars have tried to predict stock prices using different data resources, different research methods, and different evaluation indicators.

In general, the analysis approaches can be classified primarily based on the input factors. The main stock analysis or prediction algorithms can be categorized into three groups: Fundamental Analysis, Technical Analysis and Analysis Using External Information (Khadjehi, Aghabozorgi, Ying, & Ngo, 2014).

2.1.1 Fundamental analysis

The Fundamental analysis follows the assumption that prices of all stocks are determined by their internal value, which corresponds to the potential profitability including company values, expected dividends and so on. Therefore, the fundamental analysis focuses on whether the stocks are overvalued or undervalued. If so, the stock prices will decrease or increase until the stock prices match the internal stock values.

In order to evaluate the internal value of the stock, the input data usually concern the Macroeconomics aspect, industries aspect and company aspect (Hu et al., 2015). Macroeconomics aspect data are usually macroeconomic indices including GDP (Gross Domestic Product), CPI (consumer price index) and M1and M2 money supply indices. Industries aspect data are correlated with industry status like industry stock indices, the situations of competing companies. Company aspect factors are related

to the evaluation of the statues and profitability of the company in question. The commonly used company aspect factors include: PCF (Price-cash flow ratio), D/BE (Debt over book equity), SG (Sales growth), TG (Turnover growth), ROA (Return on assets) and P/B (Price-book value ratio) (Contreras, Jiang, Hidalgo, & Núñez-Letamendia, 2012). Table 1 presents some previous works applying fundamental analysis.

Table 1 Research on market prediction based on fundamental analysis[1]

Reference	Input variables	Processing method	Predict target	Output performance
(Lin, P. & Chen, 2007)	39 financial variables	GP, Fuzzy system	8 electronics companies in Taiwan Stock Exchange (TSE)	Significantly better than B&H
(Contreras, Jiang, Hidalgo, & Núñez-Letamendia, 2012)	PCF, D/BE, SG, TOG, ROA, P/B		Company's stock price in the S&P 500 (113~272 companies in different years) for which data are available	Significantly better than B&H
(Chen, Chen, & Lu, 2017)	Financial news, industrial environment indicators and 16 fundamental indicators	CCR (Charnes, Cooper, and Rhodes) model	Quarterly stock price	Around 70% direction accuracy
(Shen & Tzeng, 2015)	17 attributes	Combined soft computing model based on DRSA, FCA, and DEMATEL	Models comparing via HPR	HPR is slightly higher than 0. No comparison with B&H
(FUNG, SU, & ZHU, 2010)	12 indicators	Linear regression	Correlation between the stock divergence with indicators.	The correlations are significant. No prediction provided.

[1] The abbreviation introductions in the Appendix

2.1.2 Technical analysis

The theoretical basis of technical analysis can be roughly summarized into two assumptions (Khadjeh Nassirtoussi et al., 2014). The first is that the movement of price indicators reflects all the related factors and impacts. The second is that the patterns from the past will repeat in the future. Therefore, the researchers who believe in technical analysis put their effort into generating trading rules based on historical stock market behavior and applying those rules in predicting future stock price changes.

The traditional view of technical analysis is based on the theory that “price stand for everything” (also known as ‘Dow Theory’) (Brown, Goetzmann, & Kumar, 1998), In other words, price by itself is sufficient to predict price movement. Thus, the traditional technical analysis only uses prices or indicators that directly related to stock prices, for instance, open and close price, volume, volatility, and flow-of-funds (Hu et al., 2015). In the early stage of theory development, investors have used simple, easy-calculated trading rules like Filter Rules, Moving Average and Trading Range Breakout Rules (Yu, Nartea, Gan, & Yao, 2013). In higher-level analysis, the mathematical models and pattern recognition techniques have been applied in research, such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees or Naïve Bayes (Khadjeh et al., 2014), Autoregressive integrated moving average (ARIMA) (Wang & Leu, 1996) , Fuzzy system and Genetic algorithm (GA). Table 2 presents some previous research applying technical analysis in stock prediction. In recent years, more complicated and advanced indicators related to high-frequency trading were exploited. Nousi used Limit Order Book (LOB) data to forecast the Mid-Price movement (Nousi et al., 2019). (Tran, Iosifidis, Kanniainen, & Gabbouj, 2019) applied traditional neural network models by inputting LOB data and obtained the results with great accuracy.

Table 2 Research on stock prediction based technical analysis[2]

Reference	Input indicators	Processing method	Predict target	Output performance
(Lv et al. 2019)	Trend, volatility, cash flow, CCI, RSI, ATR, TRIX	LR, SVM, CART, RF, NB, XGB, MLP, DBN, SAE, RNN, LSTM, GRU	609 stock prices in 9 different industries.	Significantly better than B&H in majority cases
(Kaucic, 2010)	MAs, ROC, K&D, MACD, OBV, EMV	PVC, BMA, BOOST	S&P 500 index	Significantly better than B&H only when the downtrend period
(Hu, Feng, et al. 2015)	7 trading-rule involve: Stock price. Volume, MA, VMA, MACD, RSV	XCS, eTrend, NN, DT	Shanghai Stock Exchange Composite Index, industrial Index and the Business Index,	Only “eTrend” predictions have good results
(Avdoulas, Bekiros, & Boubaker 2018)	Historical price	TAR family nonlinear models (SETAR, STAR, LSTAR), ARMA, GA	5 stock market indexes: FTSE MIB, PSI-20, ISEQ-20, FTSE, IBEX 35	No correlations, no predict power
(Sezer, Ozbayoglu, & Dogdu, 2017)	RSI, SMA	GA, MLP	30 stocks	are similar to the BAH strategy. No significant better
(Fu, Chung, & Chung, 2013)	Buy/sell rules involve: MA, MACD, STC, RSI, W%R, MFI, MTM	Traditional GA and Hierarchical GA	8 Hong Kong stocks	Significantly better than B&H in downtrend period. Significantly worse in the uptrend period.

[2] Abbreviation introductions in Appendix

Reference	Input indicators	Processing method	Predict target	Output performance
(Yudong & Lenan, 2009)	Historical prices	IBCO, BPNN	Stock price in 1 day and 15 days	worse than B&H
(Nousi et al., 2019)	LOB data matrix	SVM, SLFN, MLP	The mid-price movements of 5 Finnish stocks	No trading decision made. Only models comparison.
(Tran et al., 2019)	LOB data matrix	SLFN, LDA, MDA, MTR, MCSDA, BoF, NBoF, SVM, CNN, LSTM	The mid-price movements of 5 Finnish stocks	Some models obtained significantly high accuracy in direction prediction.

2.1.3 Analysis using external information

Since the development of Web 2.0 and the invention of social media, a huge amount of online data has become available for stock prediction researchers. Many scholars tried to predict the financial markets using data that are not directly related to stock prices or markets such as online financial news, newly published policies, company emergencies, visits numbers of the company website and social media information.

Analysis using external information is a new area and there is no clear and consistent taxonomy about it. Some scholars classified it into technical analysis because it borrows the idea that using the historical pattern (although it is not price movement pattern) to predict future behavior (Hu et al., 2015). However, in this thesis, the research that using external information is categorized as an individual group.

Analysis using external information refers to the financial prediction analyses using the data and factors that are outside the financial markets. It uses statistical models or machine learning algorithms to excavate the correlation or behavior pattern between external data and the internal indicators in stock markets. Different from the traditional technical analysis using continuous time-series numerical data, the external factors are usually unstructured and some addition pre-processing including data filtering, text mining and feature extracting are needed.

Based on the data resource, the processing might be different. Table 3 lists some research on stock prediction based on external information.

Table 3 Research on stock prediction based on external information[3]

Reference	External information	Processing method	Predict target	Output performance
(Feuerriegel, Ratku, & Neumann 2016)	Financial news topics	LDA, Abnormal return, Event Study	Stock abnormal returns	Only a few topics have an impact on stock abnormal returns
(Telang & Wattal 2007)	Software vulnerability announcements and software company characteristics	Event Study, normal return, abnormal return	Stock abnormal returns	Averagely, an announcement causes a 0.6% decrease in company value. Smaller or competitive company loss more
(Jeong, Lee, & Lim 2019)	IT breach Security Investment firm industry and size	Event study	Price abnormal return	Significant on Security Breach in IT-industry firm. Not significant on security investment
(Rosati et al., 2017)	breach announcements	Event study	Volume, Price, bid-ask-spread	evidence of a short-term effect of data breach announcements on the bid-ask spread and trading volume
(Cergol & Omladič, 2015)	Wikipedia page visits, Google ticker, company names search counts in Google	Linear regressions (Fama-Macbeth regression, Fama and French model)	Company stock price in the S&P 500	Evident correlation. Significantly better than B&H strategy.
(Wei & Wang, 2016)	Wikipedia page visits Historical price	RFC	Company stock prices	No significant improve from Wikipedia data
(Cooper, Gulen, & Ovtchinnikov, 2010)	Political contribute indicators (e.g. number of candidates that firms support)	Time-series regression (MKT, SMB, HML and UMD models)	Company stock price (yearly)	correlations are strongest for firms that support a greater number of candidates that hold office in the same state that the firm is based in. No simulated investment tests.

[3] Abbreviation introductions in Appendix

2.1.4 Stock predictions based on sentiment analysis

Online text data is a hidden treasure for stock prediction research. It has high volume, speedy and real-time updates, and wide diversity. However, online text data is unstructured, hard to process and not comprehensible for computers. Sentiment analysis is a simple and straightforward technique for text processing. It allows computer programs to extract the opinions or emotions from sentences or articles without fully understanding of them. In addition, the huge amount of data and the high density of data makes it possible to treat sentiment as continuous data. Thus, sentiment analysis is suitable for stock predictions using text data.

The core process of sentiment analysis is converting the text into numerical data or binary data. Multiple methods can be used to conduct the task. For instance, a machine learning classifier that was trained with manually classified text data. Successful work from Jianqiang, Xiaolin, & Xuejun (2018) who used Deep Neural Network to build up a sentiment polarity classifier for Twitter text. Another commonly used technique is using pre-defined word-to-sentiment lexicons or dictionaries like “SentiWordNet” and “Harvard-IV-4”. Yu et al. have tried to improve the lexicon performance by building a financial special lexicon (Yu, Wu, Chang, & Chu, 2013).

The majority text data resources for stock prediction include financial news, which presents the company conditions or macro-economic environment, and social media data, which reflects the opinion of the public and investors. These two data resources are discussed below.

2.1.4.1 Financial news

Financial news, being the most commonly used information resource for investors, provides articles and messages fast and consistently. It mainly reports information related to branch companies, technology developments, load capital, management changing, mergers and acquisitions, earnings reports and so on (Feuerriegel et al., 2016). The area of majority financial news can be classified into macro-economic

environment, industry situation or company condition. All of them are stock value determining factors in fundamental analysis. Therefore, financial news has become the most straightforward and feasible information resource that attracted many researchers' attention. Table 4 lists some research on stock prediction based on sentiment analysis with financial news.

Li et al. applied sentiment analysis on Hong Kong news and the Hong Kong stock market (Li, Xie, Chen, Wang, & Deng, 2014). his experiment showed that the market capacity had a strong correlation with same-day news amount, but the sentiment of news did not have predictive power for future stock movement. Strauß, Vliegenthart, & Verhoeven (2016) proposed an experiment on the relationship between Dutch news articles and the corresponding stocks by using the time-series vector autoregression method. The results showed that the finance news had no significant effect or predict power on future stock price, but the results clarified the price affected people's sentiment.

On the other side, there are successful attempts on the stock price prediction. Hagenau, Liebmann, & Neumann (2013) and Shynkevich, McGinnity, Coleman, & Belatreche (2016) applied a good feature extracting method that the appearance of each word or phrase was an individual dimension. It made the research resulted in an impressive validation accuracy. However, the results are questionable because too many dimensions might cause overlearning. In addition, they built the corpus using the all dataset which made the validation on early samples were using later training data. These two problems might lead to the unreliability of the results.

Li, et al. (2014) also provided a representative article. He tried to use the "real trend" of the price as the training label by smoothing the price vibration, which made the prediction and simulated investment performed well. In addition, a comparison between companies with different sizes and industries has also been implemented. It showed that the IT, social service and wholesale and retail trade are the most predictable industries.

Table 4 Research on stock prediction based on sentiment analysis with financial news[4]

Reference	Text Processing method	Pattern learning method	Predict target	Output performance
(Li et al., 2014)	Pre-defined sentiment dictionaries	SVM	Daily stock price	Not good
(Strauß, Vliegenthart, & Verhoeven, 2016b)	Pre-defined sentiment dictionary: Dutch version of the Linguistic Inquiry and Word Count (LIWC)	vector autoregression	Daily stock price	Not predict power from sentiment to price, but stock price influence news sentiment
(Hagenau et al., 2013)	bag-of-words, words-combination, pre-defined sentiment dictionary: Harvard-IV-4	SVM	Daily stock price in the UK and Germany	The direction accuracies and simulation profits are higher than the random guess. No comparison with B&H
(Shynkevich et al., 2016)	bag-of-words TF-IDF	SVM, kNN, MKL(Multiple Kernel Learning)	Daily stock price	Direction accuracy is good (56% - 81%). No simulate trading strategy return
(Li et al., 2014)	eMAQT, TF-IDF pre-defined dictionary (Harvard-IV-4)	relevance language model (RLM) (SVM, NB, support vector regression)	Stock trend	52-64% direction accuracy varying based on models and industries. Significantly better than B&H strategy

[4] Abbreviation introductions in Appendix

2.1.4.2 Social media

From the pens and letters to mobile devices and webpages, the development of the Internet and communication technology brings us the convenience of communication which changes our lifestyle. Being considered as the most innovational invention (Kaplan & Haenlein, 2010), social media really plays an important and non-replaceable role in our life. Referring to Obar & Wildman (2015) and Kaplan & Haenlein (2010), social media can be defined as follows. Social media is a kind of Web2.0 Internet-based application that consists of public user-generated content. It also provides users and groups created user-specific profiles and connects them together.

Nowadays, social media has a very large number of users and high activity. Almost every groups and companies use social media as one of the most important channels to make announcements and advertisements. These characteristics make social media become the best way to monitor and research public-related topics such as public emotions and information propagation. For instance, Imran, Castillo, Diaz, & Vieweg (2015) provided a survey about using social media data to research the impact of emergency events and Ashley & Tuten (2015) researched how to use social media to help marketing.

Social media text is another important data source for stock prediction research. Comparing with traditional media, social media provides a larger amount of data, more content creators and faster propagation. However, on the other hand, it is more noisy, harder to collect and process.

Different from financial news, the majority text in social media does not provide any new information that directly impacts the market. However, it impacts the market by sharing and spreading users' emotions and opinions which are likely to be exaggerated and irrational (Lerman & Ghosh, 2010).

Table 5 presents some previous research that attempted to predict the stock movement using social media text. Nguyen et al. (2015) implemented an aspect-based sentiment method and SVM method on Yahoo! Finance Board and got 54% average

accuracy on the predictions of price movement direction, which was not good enough to confirm the relationship and therefore cannot profit from the predictions as well. Bollen & Mao (2011) researched how emotions of the whole society influenced the DJIA index movement. In this research, the authors used “OpinionFinder” to rate every single twitter in six mood categories (Calm, Alert, Sure, Vital, Kind, and Happy). The experiment showed that only the feature ‘calm’ had a significant correlation with stock price change in three-day lag. Sul, Dennis, & Yuan (2014) extended Bollen and Mao’s research into predicting specific stock price instead of the overall index. The authors used a linear regression model and found that the sentiment and the stock price change on the same day had a significant relationship. This research also showed that sentiment can help predict the possible stock price in 10 days, but not what it will be tomorrow. Yu, Duan, & Cao, (2013) examined the prediction power of the text on Twitter, Google Blogs, BoardReader, and financial news at the same time. Naïve Bayesian text classifier (NBTC) was applied to evaluate the sentiment in the text, where the NBTC is based on the likelihood that each word presents in positive or negative texts. After being processed by the Fama-French momentum four-factor model, the result showed that the sentiment had no significant correlation with the absolute return of price, but the correlation between the absolute return of price and the total number of news or social media posts were significant. Siikanen et al. (2018) introduced another alternative of stock market analysis. Instead of directly predicting the stock price change, this study analyzed the stock holding willing of the passive households. The social media data were processed by Social Data Analytics Tools (SODATO), and the data analysis methods were linear regression. This study showed a significant correlation between the company-related social media posts and the stock holding willing.

Table 5 Research on stock prediction based on sentiment analysis with social media data[5]

Reference	Data resource	Text Processing method	Pattern learning method	Predict target	Output performance
(Nguyen et al., 2015)	Yahoo Finance Message Board	TF-IDF, LDA, JST model Pre-defined dictionary (SentiWordNet)	SVM	Daily return of 18 stocks	2% accuracy higher than the traditional method. 54% averagely.
(Bollen et al., 2011)	Twitter (tweets without filter by topics)	Pre-defined dictionaries or packages ('OpinionFinder OP' and 'Google-profile of Mood States GPMS')	Granger causality analysis, AR, Self-organizing Fuzzy Neural Network SOFNN	Dow Jones Industrial Average	Good direction prediction accuracy. No simulated profit.
(Sul, Dennis, & Yuan, 2014)	Twitter (tweets that are talking about the specific company)	Pre-defined dictionary(Harvard-IV)	Cumulated Abnormal Return	Daily stock price	Significant relationship on the same day and 10 days lag
(Yu et al., 2013)	1. Blog (Google blogs). 2. Forum (BoardReader) 3. MicroBlog (Twitter). 4. Conventional news 5. Technical analysis indicators	Naïve Bayesian text classifiers	Fama–French momentum four-factor model abnormal return	Daily stock price	only numbers of tweets and news have significant correlations with the price. No prediction or simulated invest are implemented
(Siikanen et al., 2018)	Facebook	SODATO	Linear Regression	Stock holding amounts of investors	Significant correlated. No prediction made.

[5] Abbreviation introductions in Appendix

2.1.4.3 Sentiment analysis applications in another field

Besides the stock prediction, sentiment analysis can be applied in many other fields. The success of some previous research confirmed that sentiment analysis is valuable and useful.

Bai (2011) researched the sentiment analysis performance in different text types including movie reviews, financial news, NBA (National Basketball Association) news. The model 'Markov blanket classifier' was applied in sentiment analysis.

Product or service review analysis is a common research topic. Siering, Deokar, & Janze (2018) carried out remarkable work. The authors used sentiment analysis to find which specific aspects of service in aircraft made customers most willing to recommend the aircraft company. The sentiment factor for each review came from the ratio of positive words and negative words in the review sentences. Two analysis methods were applied in the research: The Descriptive statistics analysis and Prediction analysis. Specifically speaking, the linear correlation P-value test and the accuracy analysis of SVM cross-validation has been implemented. The results showed that sentiment analysis can significantly improve text review utilization and help airline companies understand the customer better.

Desmet & Hoste (2013) applied sentiment analysis method in suicide notes in order to detect what specific emotions could lead to suicide. The authors manually classified related words into 16 emotions categories and analyzed the correlation between aspect emotion frequency and suicide. Ceron, Curini, & Iacus (2015) used sentiment analysis to monitor the American and Italian Electoral Campaigns.

2.2 Economic explanation about market predictability.

While mathematicians and computer engineers applied different methods in making predictions and profits, the stock market also attracted economists who looked for the new theoretical explanations or modifying the existing theories for market predictability.

As a member of markets, stock markets obey general market theories and rules. However, different from traditional markets, people trade investment opportunities instead of real items on stock markets. Therefore, stock markets have some special features that make it distinguished from other markets. Many scholars researched stock markets from different aspects, market predictability is one of them. The most representative theories and schools are discussed below.

2.2.1 Price theory

Price theory is the most well-known theory and the most basic idea in Economics, it says that the price in a market depends on the supply and demand in that market (Mankiw, 1998). Furthermore, the price will change alongside the change of supply or demand. Specifically say, the demand and supply in stock markets respectively correspond to optimistic investors (the investor who believes the stock price is going up) and pessimistic investors (the investor who believes the stock price is going down). Once the numbers of optimistic or pessimistic investors are known, stock price change would be predictable. In this work, the sentiment of Reddit users can be considered as the sample of all investors, so the ratio of the optimistic and pessimistic user in Reddit can represent the ratio of all optimistic investors and pessimistic investors to a certain degree. Based on the sampling theory, although it would not be completely accurate, it still helps to predict the real ratio and stock price.

2.2.2 Behavioral economics

Social media is not only a place for people to present their opinion but also a place to receive information at the same time. It means that the text on social media not only reflects posters' opinions and expectations, but it also provides an impact on the market (Khadjeh et al., 2014; Robertson, Geva, & Wolff, 2006).

In the field of behavioral economics, scholars found that investors' decision making is not fully based on objective information, it is affected by investors' mood and emotions (De Long, Shleifer, Summers, & Waldmann, 1990). The previous work from Nofsinger

(Nofsinger, 2005) introduced that social interaction affects people's financial decisions. Therefore, it is reasonable to believe that the text on social media will influence other investors' emotions and decisions. In addition, the emotion itself, apart from the forecasting and opinion that is revealed from it, has an impact on the investors' financial decisions (Bollen et al., 2011; Lee & Andrade, 2015).

2.2.3 Efficient Market Hypothesis and Adaptive Market Hypothesis

The debate on whether the market is predictable has lasted for many years. In 1965, (Fama, 1965) proposed his Efficient Market Hypothesis (EMH). According to the theory, markets with well information access will immediately auto self-adjust against any impact, which results in unpredictable markets. In 1970, (Fama, 1970) pushed his theory forward by categorizing the markets into three classes based on the information available, the more information the market involvers have, the more 'efficient' the market, which means, more unpredictable it is. Representing the other side of the controversy, behavior economists believe that human is not always rational in capital markets (Oberlechner & Hocking, 2004), different investors have different weights and bias on information resources. The irrationality and bias of investors are one of the reasons that markets are theoretically predictable. In 2005, (Lo, 2005) came up with the Adaptive Market Hypothesis (AMH) which combined the EMH with behavioral economics, it gave a clear explanation about how 'efficient' works in stock markets and confirmed that stock market is predictable in a certain condition. Urquhart & Hudson (2013) examined AMH in US, UK and Japanese stock markets with very long-run data. The examination showed that the AMH had a better description of the stock market behavior than EMH. This result proves that the markets are not always efficient, and the trading strategies have the potential to be profitable based on environment conductivity.

Based on the three theories or hypotheses above, stock markets have a probability to be predictable and it is worth to implement further research on it.

3 RESEARCH METHODS

In this chapter, the methods, algorithms, and models that have been used in this thesis are illustrated. The methods are basically the text processing methods including the sentiment analysis and some basic text preprocessing and cleaning techniques, and selected machine learning algorithms.

3.1 Text mining and sentiment analysis

Ronen Feldman and James Sanger gave a definition of text mining in 2007:

“computational linguistics research that transforms raw, unconstructed, original-format content into a carefully structured, intermediate data format. Knowledge discovery operations, in turn, are operated against this specially structured intermediate representation of the original document collection.”(Feldman & Sanger, 2007)

Text mining is regarded as extracting information from text data to help future analysis and decision making. Although there is no clear categorizing of the text mining tasks, the text mining tasks mainly include data retrieval, text categorization, text clustering, text summarization, information extraction, and sentiment analysis.

3.1.1 Sentiment analysis process

Sentiment analysis (also known as ‘opinion mining’) is defined as finding the opinions or attitudes of authors about specific entities (Feldman & Sanger, 2007). Sentiment analysis is one of the hottest topics in text mining because it can be applied to many data domains, especially social media. Sentiment analysis techniques offer organizations, market managers or customers a possibility to monitor their reputation and feedback so that they can act accordingly.

In this thesis, the sentiment analysis focus on the polarity sentiment (positive or negative) contained in each post or comment in Reddit which can be used to represent

an estimation of the authors' (Reddit users') opinion or forecast about the stock price change.

The sentiment analysis method in this thesis works as shown in Figure 1.

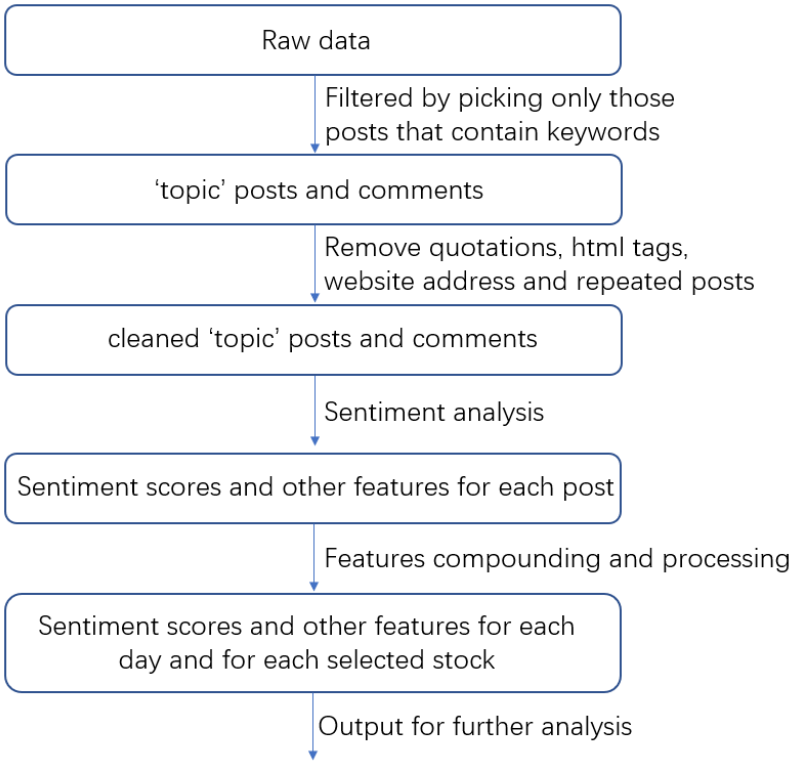


Figure 1 Process of Sentiment analysis

The text mining process starts with the filtering process. In this step, the posts that contain the company-related keywords are selected. The keywords contain companies' full name, commonly used nicknames and the Stock Symbol in the stock market. For example, the keywords for the company 'Alphabet Inc.' are 'Alphabet', 'Google' and 'GOOG'. Besides filtering, a cleaning process is also applied to the database. All the quotations, advertisements, HTML tags, website links, and repeated posts are removed by keywords or by manually. After that, the selected posts are analyzed by applying two pre-defined sentiment analysis packages: 'VADER' (Hutto & Gilbert, 2014) and 'SentiWordNet' (Baccianella, Esuli, & Sebastiani, 2010) (The details about these sentiment analysis packages will be introduced in following sections). The purpose of this process is to transform text data of posts or comments into numerical scores representing the sentiment and attitude of the authors. Last is the features

compounding process. In this step, the features of each day are calculated based on compounding the features of all posts or comments that created on that day. Posts are first sorted into each day by their created date. The scores of the posts created in the same day are used to calculate the sentiment scores and other features for that day (The feature extraction process will be introduced in Chapter 5).

3.1.2 VADER sentiment package

VADER sentiment package is an open-source rule-based sentiment analysis package and lexicon. It is invented especially for text data on social media. The experiments showed that it has a very good performance on classifying tweets (posts on Twitter), movie reviews and other types of online text (Hutto & Gilbert, 2014).

The majority of sentiment analysis approaches can be divided into polarity-based, where sentences are only classified into positive and negative, or valence-based, where the intensity of sentiment (usually numerical score) is also presented. VADER is valence-based that it shows the difference between slightly positive and strong positive sentences, which, in this paper, corresponds to the investors with strong confidence about the stock price and the investors only have little interesting in purchasing the stock.

Since VADER is designed for analyzing the sentiment score for sentence, a function ‘sent_tokenize’ from NLTK (natural language toolkit) (Bird, Klein, & Loper, 2009) is applied to separate the posts or comments into sentences, which would be processed individually later. NLTK is a Python opensource platform package for building Python scripts to work with human language text data. It provides multiple text-processing functions, for instance, tokenization, stemming and tagging. In addition, it also provides direct access to many other text processing packages, like VADER that used in this thesis.

The analysis process of VADER is shown in Figure 2.

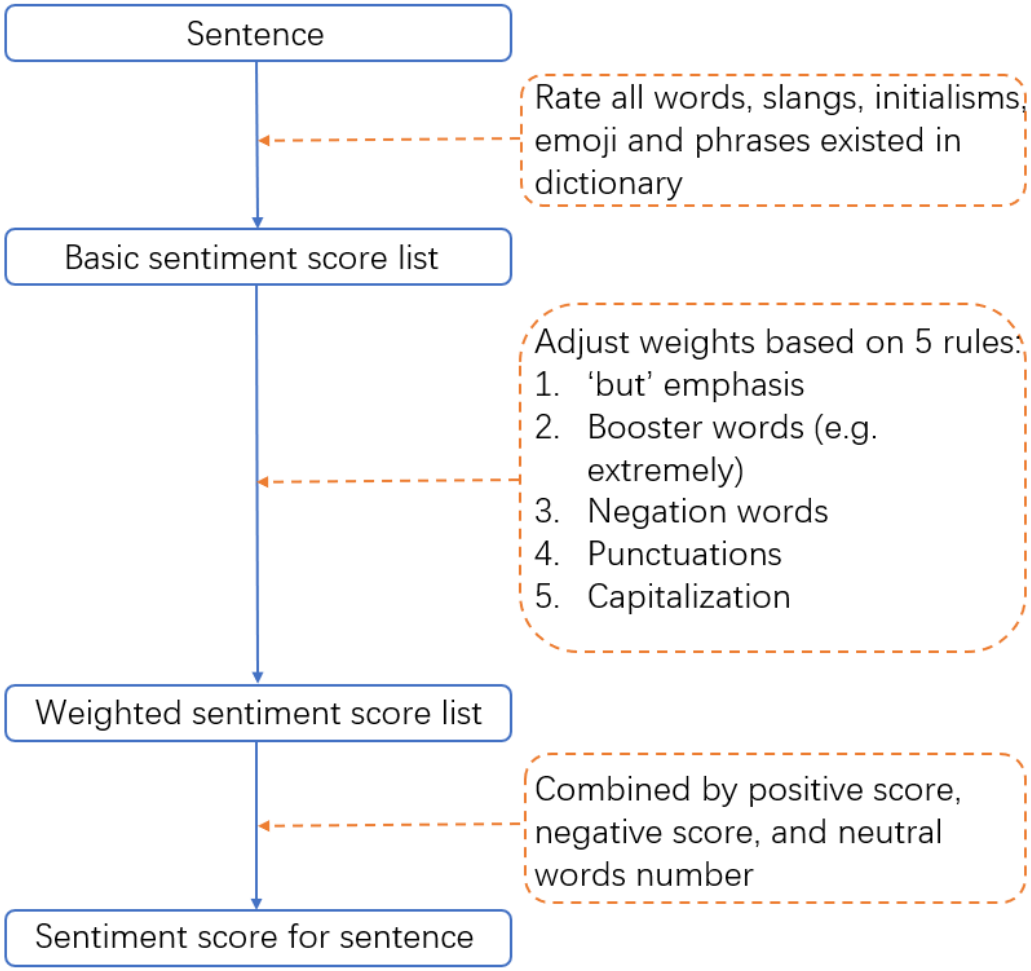


Figure 2 Process of VADER sentiment package

When VADER sentiment package is used to analyze a sentence, it first checks whether words, slangs, initialisms, emoji (in 'utf8') and phrases (all of them are called 'words' for convenience in follows) exist in lexicon and give them corresponding rating (from -4 to 4, representing very negative to very positive). After that, the sentiment scores of words are weighted based on 5 rules: (1) 'but' cuts the words' score before it and emphasis the words' score after it; (2) Booster words, like 'extremely' and 'barely', impact sentiment score of next word by adding a weight parameter; (3) Punctuations. For instance, the Exclamation mark ("!") carries more intense sentiment than the Period("."). The sentiment score of the whole sentence is emphasized based on its polarity; (4) Negation flips the polarity of the sentence; (5) The sentiment scores on capitalization words are emphasized. Finally, a set of sentiment scores for the sentence is outputted. The scores for the sentence are calculated as below:

$$\text{positive\_score} = \frac{\text{sum of scores of all positive words}}{\text{sum of scores of all words + word count}} \quad (1)$$

$$\text{negative\_score} = \frac{\text{sum of scores of all negative words}}{\text{sum of scores of all words + word count}} \quad (2)$$

$$\text{neutral\_score} = \frac{\text{neutral word count}}{\text{sum of scores of all words + word count}} \quad (3)$$

$$\text{compound\_score} = \frac{\text{sum of scores of all words}}{\sqrt{(\text{sum of scores of all words})^2 + 15}} \quad (4)$$

In this thesis, the compound score is used as the sentiment score of the sentence.

In addition, an example of applying VADER into a sentence from social media posts is shown in Figure 3.

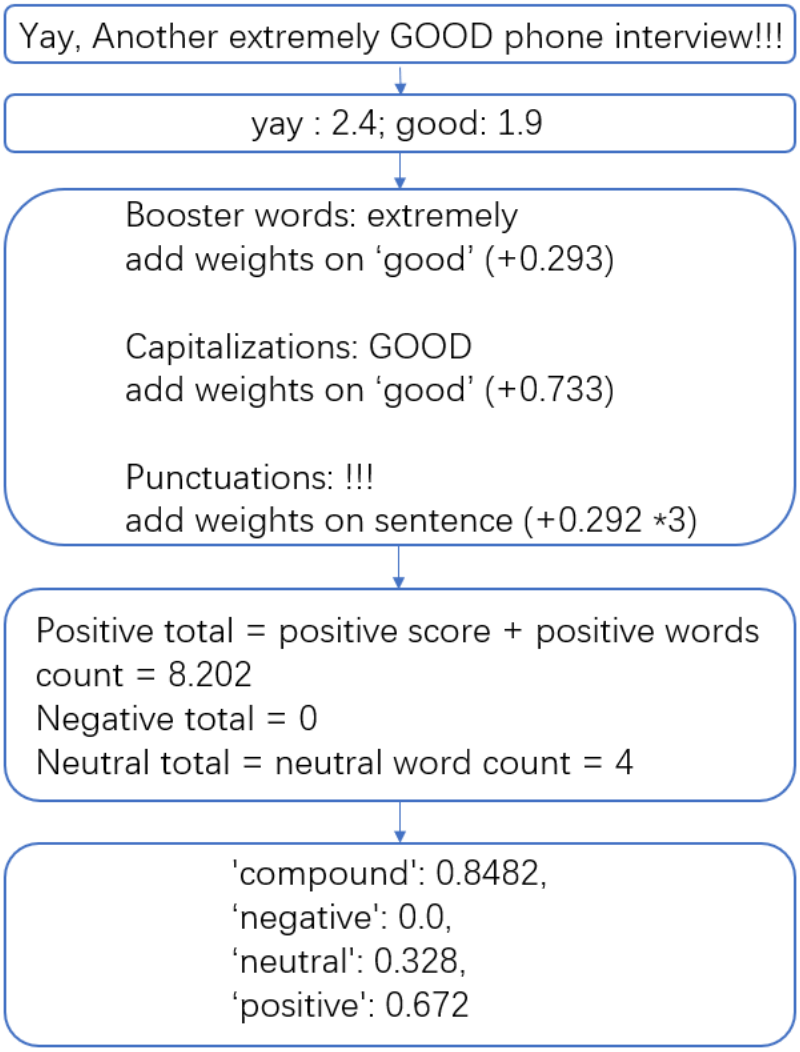


Figure 3 Example of sentiment analysis with VADER package

After processing all sentences' sentiment, the compound sentiment scores for posts or replies are calculated as function below:

$$S_{overall} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} abs(S_i)} \quad (5)$$

Where S_i is the sentiment score for the $i-th$ sentence in the post or comment. $S_i > 0$ if the sentence is positive, $S_i = 0$ if neutral and $S_i < 0$ if it is negative.

3.1.3 SentiWordNet

SentiWordNet (SWN) (Baccianella et al., 2010) is a lexicon resource devised for supporting sentiment analysis and classification. It is a dictionary that assigns 3-dimension sentiment score (positive, negative and objective) to every synsets and format of words (e.g. word 'black' can be adjective or noun, which carries different sentiment). Same as the VADER sentiment package, SWN is also a valence-based dictionary.

Because SWN is only a dictionary, in order to make use of it, other tools like NLTK and corpus like WordNet (WN) (Miller, 1995) are needed. WN is a large lexical database of English which groups nouns, adjectives, adverbs, verbs into sets of synsets. It is used to identify whether the words are nouns, verbs, adjectives or adverbs in this thesis.

The processing of analysis is introduced in follows. Firstly, the sentences are separated into words by using the tokenization function from NLTK. Secondly, words will be lemmatized and tagged with their part of speech, in this thesis, only nouns, verbs, adjectives and adverbs are used. After that, the corresponding sentiment scores for every word-tag pairs will be found from SWN lexical dictionary. Finally, scores of words are added up into the final positive and negative score for sentence (although the package also provides an objective score, it does not be used in this thesis). The compound score for a sentence is calculated as below.

$$compound = \frac{S_P - S_N}{S_P + S_N} \quad (6)$$

Where S_P is the sum of positive scores of all words in the sentence and S_N is the absolute value of the sum of negative scores of all words in the sentence.

An example of SWN sentiment analysis is shown in the figure below (Figure 4).

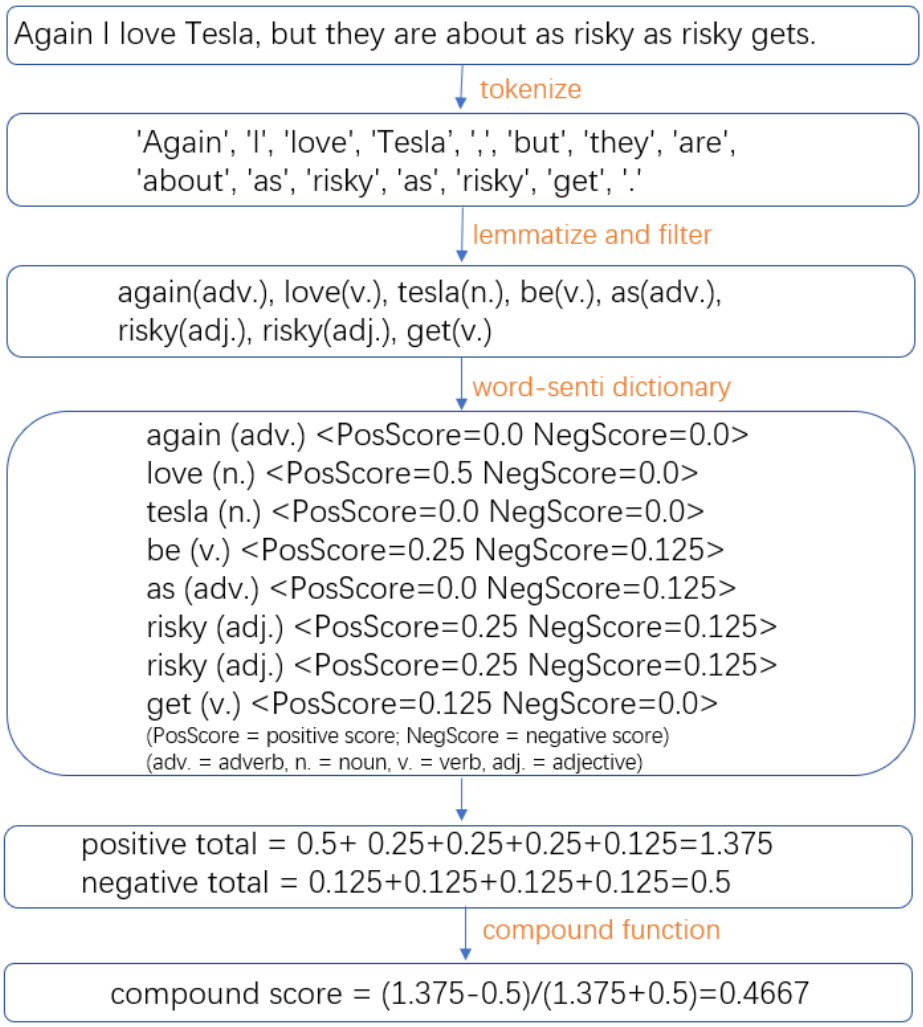


Figure 4 Example of sentiment analysis using SentiWordNet package

Although based on the experiments from other research (Hutto & Gilbert, 2014), the SWN performs worse than VADER in some situations, but it is still used to provide an alternative or one extra dimension in sentiment score estimation.

3.2 Machine Learning models

Today is called the age of “BIG DATA”, countless data and data-related technologies are created or invented constantly. Machine Learning is the most famous research

field with hundreds or thousands of various developed methods and models which have been applied to lots of research. Alpaydin (2014) gave the definition to Machine Leaning as below.

“Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data or both.” (Alpaydin, 2014)

A common task of Machine Learning is classification, which is also the task in this thesis. The Machine Learning classification task is the process of constructing a classifier using human-designed rules and historical data. Where the human-designed rules include statistical models and performance measures, the construction processing is that computers optimizing the model parameters to attain the best or approximative best values of the performance measures. Furthermore, the statistical models are functions that take the known data (e.g. historical data) as input value and output the class it belongs to. The performance measures are various, for example, the distance between boundaries, or simply the classification accuracy. The historical data are usually given in the form of input data that have variable numbers of features and a label (or multiple labels in some case) representing the category (or categories) the sample belongs to.

Although Machine Learning is commonly used as a classifier or predictor, it can also be used to analyze the relationship between variables. Based on the algorithm used, machine learning might be able to find non-linear or other more complex relationship between variables which are hard to be found by human eyes or traditional statistics methods. It called predictive analysis (further discussion in Chapter 5 and 6.2).

In this thesis, 6 machine learning algorithms are used: K-Nearest Neighbor Classifier, Support Vector Machine, Logistic Regression Classifier, Decision Tree, Random

Forest Classifier and Gradient Boosting Model. They are based on different ideas and algorithms and have their own advantages and characteristics.

3.2.1 K-Nearest neighbor classifier

The nearest neighbor classifier is one of the simplest classifiers. The object is classified by majority votes of the training samples which have top-K shortest distances to it according to the input features of the object and the corresponding input features of the training samples.

The Euclidean distance function is the most widely used distance function, which is also the method used in this thesis:

$$\mathrm{d(a,b)} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{7}$$

Where *a*, *b* are the objects, $\mathrm{a_i}$ and $\mathrm{b_i}$ are their features. *n* is the number of features for each object. d means distance.

Votes of training samples can also be modified by adding weight to voting. The weight is based on, for example, the distance between each sample with the object, the samples with shorter distance contribute more to voting. 'Inverse Distance' is applied in this thesis. 'Inverse Distance' works as below.

$$C(y) = arg\ \max_{k}(\sum_{x_i \in C_k} \frac{1}{d(y,x_i)}) \tag{8}$$

Where y is the predicting object, $C(y)$ is the predicted class that the object belongs to, x_i are samples and C_k are the sets of samples in each class *k*.

3.2.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm invented by Cortes & Vapnik (1995). Its original idea is to find an optimal line (or hyperplane in high dimension dataset) between two classes so that the margin level is maximum. Figure 5 (Scikit-learn, 2019) is an example of SVM. The function of SVM is shown as follows.

$$\max_{\boldsymbol{w},b,\|\boldsymbol{w}\|=1} M, \text{ subject to } y_i(\boldsymbol{x}_i^T\boldsymbol{w}+b) \geq M, \text{ for } i = 1,2,\dots,N \qquad (9)$$

Where M corresponds to the distance between the 'edge points' (margin level). Matrix $\boldsymbol{x}_\mathrm{i}$ is the feature matrix of the *i-th* object and variable y_i is -1 or 1 representing the class that *i-th* object belongs to. Matrix $\mathbf{w}$ and is the parameter matrix and *b* is a parameter variable that is calculated during the process of optimizing the margin level.

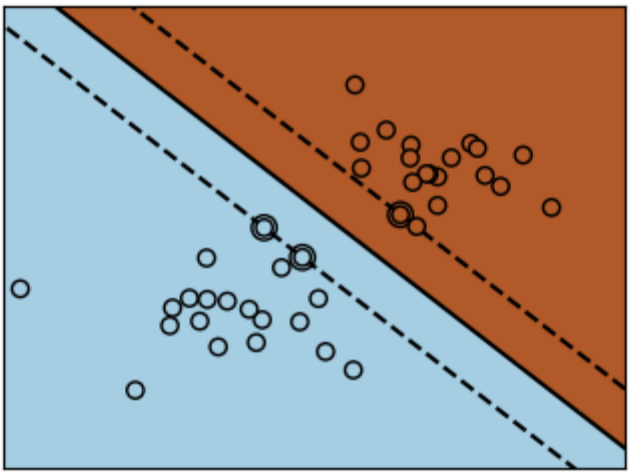


Figure 5 Example of Support Vector Machine

If the dataset cannot be well separated, the line is constructed to minimize the value in the hinge loss function (see function below).

$$\min_{\boldsymbol{w},b} \sum_{i=1}^{N} \left[max\Big(0,1-y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\Big)\right] + C\|\boldsymbol{w}\|^2 \qquad (10)$$

Where $\boldsymbol{w}$ is the normal vector to the line, $\boldsymbol{x}_i$ is the feature matrix of the i-th object. y_i is either -1 or 1 indicating the class that the i-th object belongs to. C is a constant parameter that determines the trade-off between margin level distance and the wrong-side penalty.

SVM can be extended to nonlinear classification by applying kernel trick. In the kernel trick, each inner product is replaced by a kernel function. The parameters are also reformulated to optimize the margin level or hinge loss function for inner product values instead of original input features. Essentially, the input objects are mapped to a set of

inner product values. Where the linear boundary in inner product value calculation corresponds to the non-linear boundary in original objects.

For instance, the Polynomial kernel (inhomogeneous) works as below:

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} \cdot \overrightarrow{x_j} + 1)^d \quad (11)$$

Where x_i is the sample coordinate and d is the degree of the polynomial.

3.2.3 Logistic Regression

The Logistic Regression (LR) is called a regression method but it returns probabilities of binary outcomes instead of predictions of a continuous-valued regression output variable like other regression analyses. In this way, LR is suitable for the Machine Learning classification task.

For the normal 2-classes cases, the algorithm works as below.

Firstly, all the input objects are assumed to follow the probability distribution below.

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{1+e^{-f(x|\boldsymbol{\theta})}} \quad (12)$$

$$f(\boldsymbol{X}|\boldsymbol{\theta}) = x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_{\mathrm{n}} x_{\mathrm{n}} + \varepsilon \quad (13)$$

Where $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the feature matrix and class matrix of input objects. $\boldsymbol{\theta}$ is the weight parameter matrix.

Secondly, the weight matrix $\boldsymbol{\theta}$ is trained by the maximum likelihood method. Or in other words, the weights are the maximum likelihood estimations (MLE) of the distribution of the training data. The conditional likelihood of output classes of the samples when given their input features is:

$$L(\theta|\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{(1-y_i)} \quad (14)$$

Where $y_i = 1 \text{ or } 0$ depends on the class that the i-*th* sample belongs to. x_i is the feature vector of the i-*th* sample.

In order to find the maximum of the likelihood above. As solving normal MLE questions, after applying a differentiating to log-likelihood (the logarithm of likelihood equation), what comes out is (Cosma, 2019) :

$$\boldsymbol{l}(\theta; \boldsymbol{x}) = logL(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} y_i log\ p(x_i) + (1 - y_i) log(1 - p(x_i)) \quad (15)$$

$$\frac{\partial \boldsymbol{l}}{\partial \theta_j} = -\sum_{i=1}^{n} \frac{1}{1+e^{\theta_0 + \boldsymbol{x_i}\boldsymbol{\theta}}} e^{\theta_0 + \boldsymbol{x_i}\boldsymbol{\theta}} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \quad (16)$$

The derivative formula above is not feasible to solve it accurately by setting this expression to zero and solving. However, in practice, scholars usually approximately solve it. There are lots of methods for finding the approximate optimization point. This thesis used the 'Trust Region Newton Algorithm' (Lin, Weng, & Keerthi, 2007). The algorithm starts with a randomly initialed $\boldsymbol{\theta}$, and adjust $\boldsymbol{\theta}$ towards the direction with a negative gradient. In each iteration, the adjustment is:

$$\theta_{t+1} = \theta_t - \epsilon \cdot \boldsymbol{l}'(\theta_t) \quad (17)$$

Where $\boldsymbol{l}'(\theta_t)$ is the gradient vector of θ_t and ϵ is a constant parameter called 'learning rate'.

The iteration process ends until the adjustment having no improvement or reaching the maximum iteration limit. *Figure 6* shows a simple example of an optimization path and iteration processing under a One-dimension Logistic Regression task.

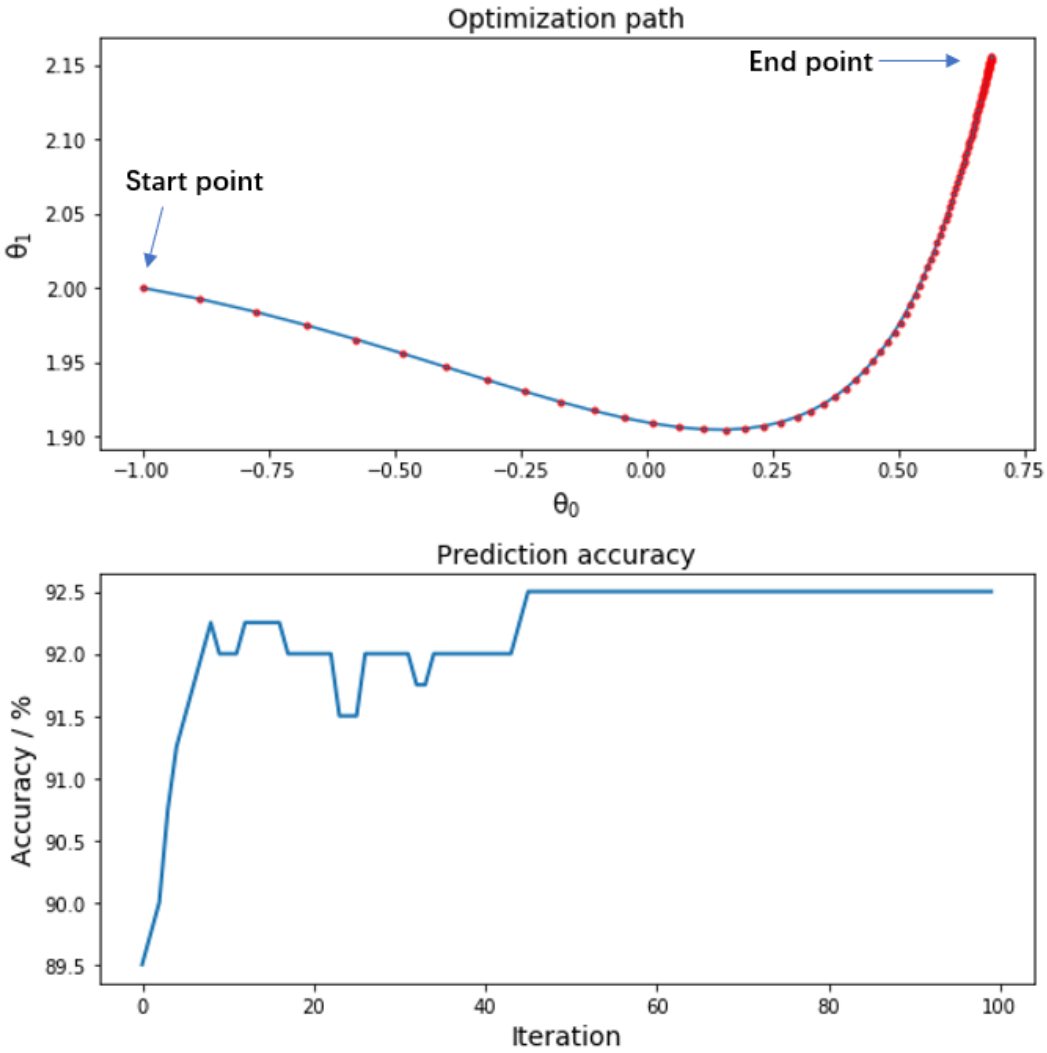


Figure 6 Example of approximately maximizing path

In the top figure, each dot corresponds to the parameter vector in each iteration. The bottom figure shows how the accuracy, or in other words, the performance of the model, is increasing along with the iteration.

After that, the training samples are projected into a one-dimension space by multiplying the learned weight matrix $\boldsymbol{\theta}$. The results are passed into the 'logistic sigmoid function' (Figure 7) to be converted into the probabilities of each class that the predicting objects belong to.

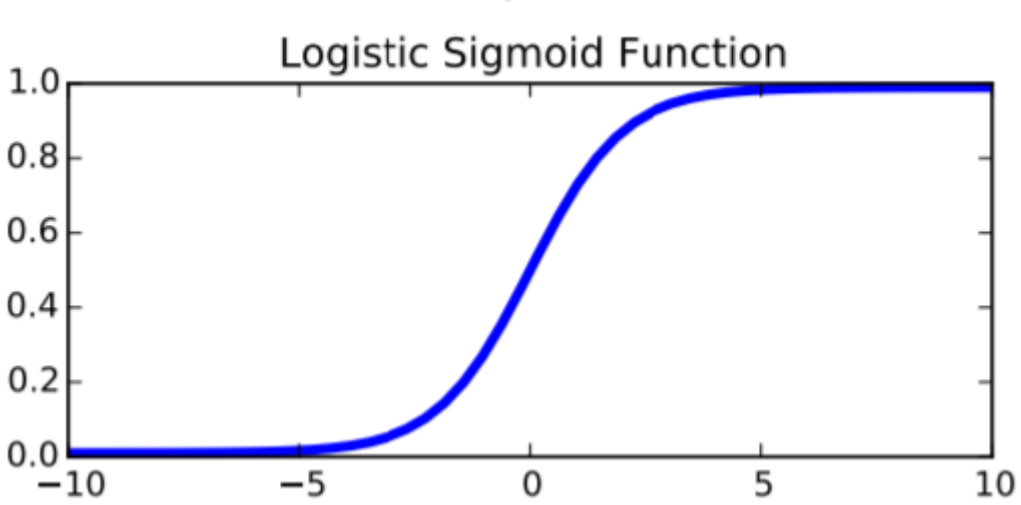


Figure 7 Logistic Sigmoid Function Figure

The final expressions are shown below.

$$P(c = 1|\boldsymbol{x}) = \frac{1}{1+exp(-(\boldsymbol{\theta}^T\boldsymbol{x}))} \quad (18)$$

$$P(c = 0|\boldsymbol{x}) = 1 - P(c = 1|\boldsymbol{x}) \quad (19)$$

Where the $\boldsymbol{x}$ is the sample feature vector and $\boldsymbol{\theta}$ is the learned parameters matrix.

Finally, the predict results of input data are the classes with higher probabilities.

3.2.4 Decision tree and Random forest classifier

The decision tree classifier is a simple but commonly used machine learning algorithm. The algorithm separates the data set (initial dataset or subset from the last node) into smaller subsets until the subsets are small enough. The final subsets called ‘leaves’ which are used to determine the class that samples in the subsets belong to.

An example figure about using the decision tree to decide whether to play tennis based on weather conditions is shown in Figure 8.

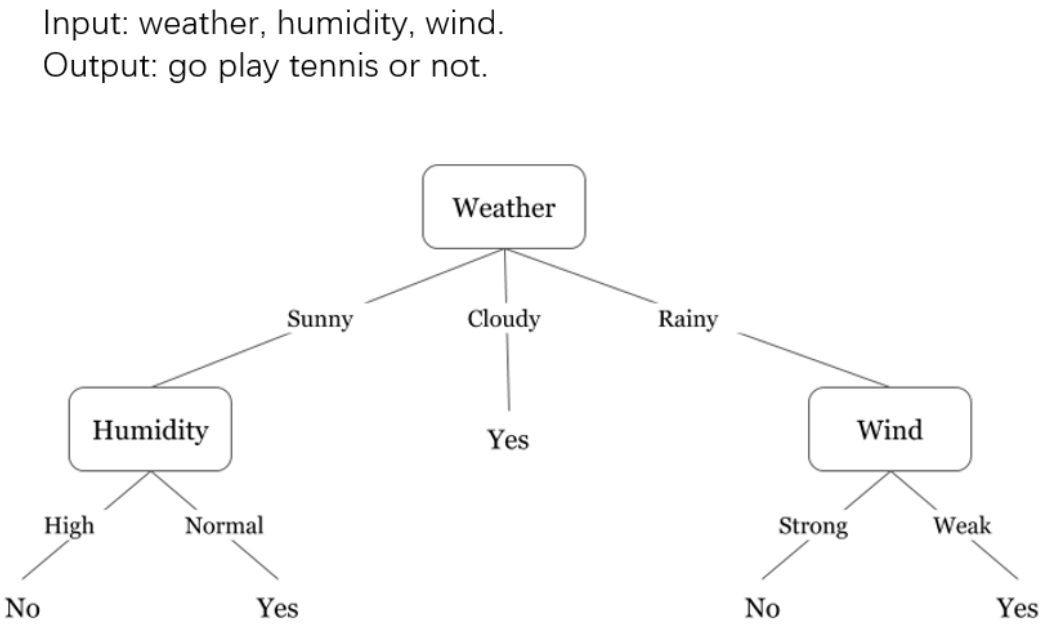


Figure 8 Decision Tree example

Being widely used in building decision trees, the 'Entropy' method aims to decide which feature should be chosen to be used in splitting on each step. It also helps to decide how many subsets are used and which boundary of continuous-valued features are used for the splitting. The 'Entropy' method sorts feature based on their 'information gain'. Which are the entropy reductions during the splitting while the entropy of the class labels in a subset of data is defined as below.

$$H(T) = -\sum_{i=1}^{J} p(x_i) log_2 p(x_i) \tag{20}$$

Where $p(x_i)$ is the percentage of each class in the subset.

Therefore, Information gain is defined as:

$$IG(T, \boldsymbol{a}) = H(T) - H(T|\boldsymbol{a}) = -\sum_{i=1}^{J} p(x_i) log_2 p(x_i) - \sum_{\boldsymbol{a}} p(a) \sum_{i=1}^{J} p(x_i|a) log_2 p(x_i|a) \tag{21}$$

Where $\boldsymbol{a}$ are the children subsets after split and $p(a)$ are their proportions (in terms of the number of samples) of the original subset.

The feature that provides more information gain would be used first.

The Random Forest Classifier (RFC) is an ensemble machine learning algorithm by establishing multiple decision trees in training and use the majority trees' voting result as the output. RFC is popular in the data mining task because it can avoid overfitting by using Bootstrap aggregating processing. In Bootstrap aggregation, training data matrix $X \in R^{m \cdot n}$ (m samples with n features) are generated into several subsets by sampling uniformly along rows and columns (samples are selected with replacement while features not). Each subset is used to fit an individual decision tree, and the trees are combined by voting for the final prediction. Different from the decision tree model, RFC usually performs better because it decreases the variance of the model and avoids the disrupting from outliers or noise.

In addition, RFC can also assess feature importance by comparing the accuracy drops between the subsets with feature absent or present. For example, losing an important feature will cause the accuracy to drop a lot, and shuffling a noisy feature have no significant impact on the result.

3.2.5 Gradient boosting model

The gradient boosting model is a classifier with forwarding stage-wise fashion. In each stage, a base learner is fitted to approximate the residuals between the designed output values and last iteration estimations by minimizing the loss function.

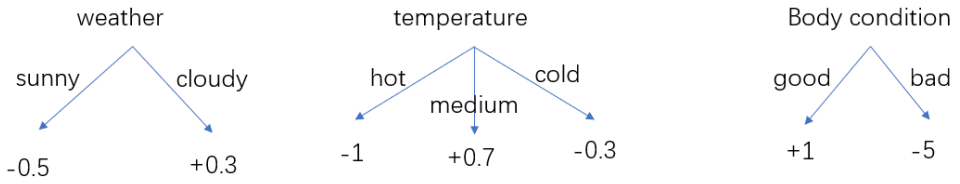
$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + arg\min_{h_m}[\sum_{i=1}^{n} L(y_i, F_{m-1}(\boldsymbol{x_i}) + h_m(\boldsymbol{x_i}))] \qquad (22)$$

Where F_m is the target estimation function of the m-th iteration, h_m is the base learner and L is the loss function. $\boldsymbol{x}$ is the input data matrix, while $\boldsymbol{x_i}$ and y_i are the feature vector and label of the i-th object.

In application, Gradient boosting is usually used with the regression trees method (also called classification and regression trees, CART). The CART consists of several trees, while each tree has similar decision rules as the normal decision tree, but with one

score in each leaf instead of a target label, the result is calculated by adding up the scores from each tree. CART works as Figure 9 below.

	weather	temperature	Body condition
Day 1	cloudy	hot	bad
Day 2	sunny	medium	good



Thresh = 0

Day1: score = +0.3 -1 +1 =+0.3 > thresh. Output Yes
Day2: score = -0.5+0.7-5=-4.5 < thresh. Output No

Figure 9 CART tree example

4 DATA COLLECTION

The text data from Reddit online forum are used in this thesis. Reddit is a social media forum with some special characteristics like the topic organization. These characteristics of Reddit make it professional and serious. Therefore, Reddit is a great text resource for researchers and investors. This chapter illustrates why Reddit is chosen and how the Reddit data are collected and processed.

4.1 Research context

Reddit is a forum website that provides platforms for news aggregation, content sharing, and discussion. The contents are organized into different boards (called 'subreddit') by their subjects. The subreddits cover countless topics including news, movies, music, video games, food, finance, and investment. In November 2017, Reddit had more than 330 million monthly visitors, ranking as the No. 5 most visited website in America (Reddit, 2017). In Reddit, every registered user has access to any subreddits and can post, reply, vote and create new subreddit freely. Topic-categorizing is the key principle that makes Reddit outstanding. Subreddits help people find their interests and provide more focused audiences, posters, and discusser, which encourage people to share their ideas and discuss seriously and actively.

Comparing with other social media like Facebook or Twitter, Reddit has its advantages. The text data in Reddit is easier to be cleaned and processed because there are a smaller number of advertisements or robots. In addition, the language used in Reddit is in a relevantly formal way instead of using a large amount of 'emoji', abbreviations and 'social media tongue'. It makes the text easier for computer text mining. However, there are also disadvantages. The majority of texts are replies or parts of conversations. It means the structure or ideas of the sentences may be missing or hiding in previous replies. Therefore, language processing algorithms can only extract partly information from the text in some situations.

Reddit is a good text resource for researching. Lots of research used Reddit as their text data resource. For instance, Zayats and Ostendorf used Reddit text data to train a conversation model with a graph-LSTM model (Zayats & Ostendorf, 2017). Olson and Neal used it to monitor and visualize the public interests and evaluate the relationship between topics (Olson & Neal, 2015). Reddit was used to track the discussion for specific items, like cigarettes (Brett et al., 2019) or drugs (Pandrekar et al., 2018). Furthermore, Reddit also contributed to training the Question Answering AI (Alambo et al., 2019). In this thesis, the Reddit text data are used to detect the public emotions and forecasting about stock prices in the stock market.

4.2 Data collection and filtering

In this study, the data from an opensource Reddit database[6] are used. The data was collected and shared by a Reddit user called “Stuck_In_the_Matrix”. According to him, the data was grabbed and saved from the Reddit API. Each reply or post (also called ‘submission’ in Reddit) was stored as a JSON format object which contains user information, the position in comment tree, the type, content and other characteristics of the post. It has roughly 40 to 150 attributes depending on posts type and created time. However, the majority of dimensions are only created for website format or database retrieval which are needless for text mining tasks. Only several involved attributes like “Author”, “Body”, “created_UTC” and “subreddit” are used. A piece of sample data is shown in Figure 10.

[6] https://files.pushshift.io/reddit/

{"author": "daKav91", "author_flair_css_class": null, "author_flair_text": null, "body": "Not to be a pessimist, but I think AAPL won't hit a home run tomorrow. No technical DD here, just my gut feeling (people holding off buying iPhones in anticipation of new ones). I hope I eat my words tomorrow, though. ", "can_gild": true, "collapsed": false, "collapsed_reason": null, "controversiality": 0, "created_utc": 1501548406, "distinguished": null, "edited": false, "gilded": 0, "id": "dkzplc9", "is_submitter": false, "link_id": "t3_6qqlhg", "parent_id": "t1_dkzf3mn", "retrieved_on": 1503655351, "score": 6, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

{"author": "shibbypwn", "author_flair_css_class": null, "author_flair_text": null, "body": "To your point though - I really think guidance is going to be more pivotal than earnings for AAPL. So if they project massive sales for 10th Anniversary of the iPhone, that could easily be more influential than last quarter's revenue", "can_gild": true, "collapsed": false, "collapsed_reason": null, "controversiality": 0, "created_utc": 1501551810, "distinguished": null, "edited": false, "gilded": 0, "id": "dkzsbkm", "is_submitter": false, "link_id": "t3_6qqlhg", "parent_id": "t1_dkzplc9", "retrieved_on": 1503656841, "score": 5, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

{"author": "do_u_think_i_care", "author_flair_css_class": null, "author_flair_text": null, "body": "Apple has been a dying company for a while now. They don't innovate, and I see them starting to fall soon.\n", "can_gild": true, "collapsed": true, "collapsed_reason": "comment score below threshold", "controversiality": 0, "created_utc": 1501565220, "distinguished": null, "edited": false, "gilded": 0, "id": "dl019rf", "is_submitter": false, "link_id": "t3_6qqlhg", "parent_id": "t1_dkzplc9", "retrieved_on": 1503661404, "score": -6, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

{"author": "zaysev36", "author_flair_css_class": null, "author_flair_text": null, "body": "I'm long term with Apple. Undervalued stock with Constant growth, and competitive value.", "can_gild": true, "collapsed": false, "collapsed_reason": null, "controversiality": 0, "created_utc": 1501598359, "distinguished": null, "edited": false, "gilded": 0, "id": "dl0hvrr", "is_submitter": false, "link_id": "t3_6qwmks", "parent_id": "t3_6qwmks", "retrieved_on": 1503669791, "score": 5, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

{"author": "MiloGoesToTheFatFarm", "author_flair_css_class": null, "author_flair_text": null, "body": "I just don't see it. Earnings with FANG have been mixed. An article about Netflix being $20B in debt just came out. If Apple misses it could trigger a melt down rather than a melt up.\n\nI know the market isn't just tech but tech has lead this rally, so it could just as easily go the other way.\n\nAdditionally the banking sector is mired in scandal again and the auto industry is fading some more.\n\nAlso, I'm just sharing my theory on the market. I'm totally open to the possibility that I'm wrong, in which case I'll chase the rally some more.", "can_gild": true, "collapsed": false, "collapsed_reason": null, "controversiality": 0, "created_utc": 1501606358, "distinguished": null, "edited": 1501608405, "gilded": 0, "id": "dl0pd4i", "is_submitter": false, "link_id": "t3_6qx98b", "parent_id": "t3_6qx98b", "retrieved_on": 1503673539, "score": 1, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

{"author": "Chad_arbc", "author_flair_css_class": null, "author_flair_text": null, "body": "I don't even think to sell it this time. It makes sense to continue holding AAPL. ", "can_gild": true, "collapsed": false, "collapsed_reason": null, "controversiality": 0, "created_utc": 1501615360, "distinguished": null, "edited": false, "gilded": 0, "id": "dl0y8fc", "is_submitter": false, "link_id": "t3_6qwmks", "parent_id": "t3_6qwmks", "retrieved_on": 1503677951, "score": 10, "stickied": false, "subreddit": "stocks", "subreddit_id": "t5_2qjfk"}

Figure 10 Data Example

The data are filtered first based on its “subreddit”. All the stock-related posts are mainly assembled in 3 “subreddits”: “r/stocks”, “r/StockMarket” and “r/wallstreetbet”. Only the comments and submissions belonging to these “subreddits” are selected. Secondly, the posts will be selected by whether their text contains the keywords for the company (see Section 3.1.1). After filtering, the data will be used for further analysis processing.

5 DATA PROCESSING AND ANALYSIS

The key ideas of the analysis processing can be summarized as 3 core components: sentiment analysis, descriptive statistics, and predictive analysis. The first part, sentiment analysis, is the process that converts the unstructured text data into well-structured numerical data using sentiment packages. Secondly, in the descriptive statistics, the evaluation between numerical data and daily price change is implemented to find the linear correlation and help to build machine learning models. After that, in the prediction analysis part, the predictions are made based on the machine learning algorithms. The performances of the prediction are the measurements and indicators of whether the correlations exist or how significant the dependencies are. The performances of the prediction are evaluated from three aspects: the accuracy of direction prediction, the daily profit efficiency and the profit of investment simulation.

5.1 Feature extraction

The first step of the data analysis is feature extraction. In this process, features that represent the characteristics of Reddit text data are calculated for further analyses. Firstly, the text data is converted into sentiment scores for each post, after that, the sentiment scores for each day are calculated based on all posts created on that day. This processing is shown in Figure 11 below.

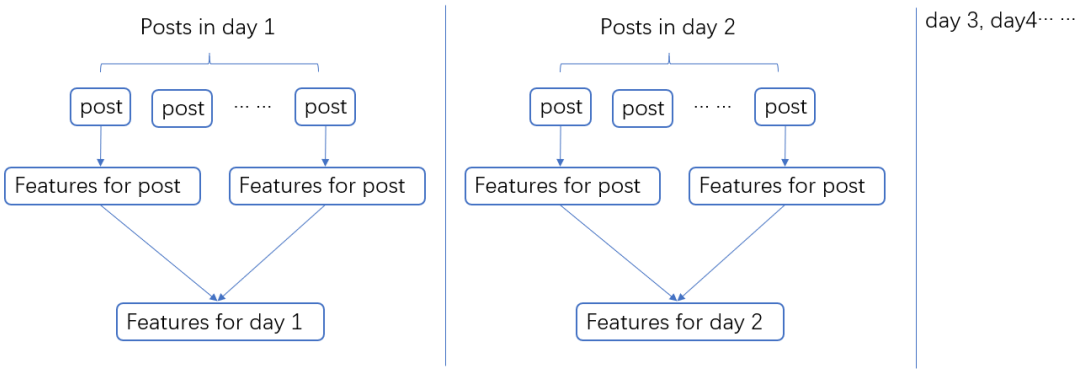


Figure 11 Feature extraction processing

5.1.1 Feature extraction for posts

Feature selection for text data is not a straightforward process. The text data contains much noise or unrelated topics and the emotions and sentiments of humans are complicated, so describing the text with only one aspect is not enough. In order to find out the features or aspects that described the text data best, several measurements were applied to make more dimensions in the feature set that describe the sentiments of the posts.

For each post, 6 sentiment scores are calculated with respect to two different sentiment packages, sentence filtering measurement, and vote score weighting.

All the sentiment features are based on the two sentiment packages. The first two features are the overall sentiment scores combined from all sentences in a post. The next two features are calculated based on the sentence filtering measurement: key sentence filtering. Although the Reddit text data are already filtered to be stock-related, people sometimes discussed under un-directly related topics like statistical analysis methods, forum rules and atmosphere, personal profit records and so on. However, the most important texts are those sentences that directly related to stock price forecasting. For instance, the sentences that authors express their opinions or forecasting about the future price change or the sentences that talk about the operating condition of companies. In this way, only the sentences that contain one or more certain keywords are counting in this sentiment score calculation measurement. The

keyword list is shown in *Table 6* below. The keywords are manually selected from top-1000 most frequent used words in all datasets.

Table 6: Keyword list

Company condition or stock features	earning, term, risk, product, products, profit, revenue, sales, profits, debt, gains, valuation, loss, profitable, overvalued, crash, sale, launch
Trade decision	buying, bought, hold, holding, selling, sold, opportunity, purchase, recommend, owned, held, purchased, throw, owning
Price forecasting	down, growth, higher, drop, grow, move, growing, fall, bullish, drops, jump, falling, increasing, predict, bull, bear, bears, bulls
Stock-related	money, price, share, shares, dividend, dividends, cost, costs, prices, call, calls, close, pick, picks, deal, returns, trends

The vote score weighting measurement imports the vote factors. As the data sample in Section 4.2, each post contains a voting score that every user can 'upvote' or 'downvote' freely. Like the content of the post is representing the author's opinion and ideas, the vote scores are representing the opinions of those users who have given the votes. Users upvote a post when users agree with or have the same idea with that post and downvoting works in the opposite direction. The posts with higher vote scores mean that the posts carrying more popular opinions. In this way, the vote score works as weights in sentiment score calculation to emphasize those widely accepted posts. The scores are calculated as below.

$$'\text{voted}'\ \text{sentiment score} = \text{sentiment score} * (1 + (\text{upvote} - \text{downvote}) * 0.2) \quad (23)$$

The 'upvote' and 'downvote' in expression corresponding to the 'score' and 'controversiality' in the dataset (as shown in Figure 10) and the 'sentiment score' in function is the whole-post sentiment score which combined from the scores of individual sentences as described earlier.

The sentiment features for a post are shown in Table 7.

Table 7 Selected Features

Features name	Feature describes
swn_overall_sentiment	Sentiment score for the whole post which processed by SentiWordNet package
vader_overall_sentiment	Sentiment score for the whole post which processed by VADER package
swn_keyword_sentiment	Sentiment score for the sentence with keywords which processed by SentiWordNet package
vader_keyword_sentiment	Sentiment score for the sentence with keywords which processed by VADER package
swn_overall_vote_sentiment	Compound of swn_overall_sentiment and vote
vader_overall_vote_sentiment	Compound of vader_overall_sentiment and vote

5.1.2 Features extract for day

The stock price is dynamically changing every second through a day. It is hard to track and analyze the real-time stock price. Instead, people usually use market-close price to represent and summarize the daily stock price changing pattern. The market-close price is one of the most important and representative price indicators of the day in investment decision making and researching as well. In order to establish a good correspondence between sentiment features and daily stock price patterns, the sentiment scores and other features are measured per day.

Since the prediction target is the market-close price, the separate time for each day is set as the market-close time (4 p.m. ET for NASDAQ stock market) for each day. It means, only the posts created before the market-close time are counted as that day's posts. This is because only the statements and discussions on the website before market close can impact the close price, and the posts after market-close time can only reflect in the next day's stock price. Conversely, users can only react to and comment about the market-close price after the market has closed.

The features for a day can be divided into sentiment-related features and activity-related features. The sentiment-related features are the means and variances of all posts' sentiment features (2*6 = 12 features in total) that created in that day and the 6 features about the activity of the forum are shown in Table 8 below.

Table 8 Features for Day

Feature names	Feature descriptions
submission_count	Total number of submissions in a day
reply_count	Total number of replies in a day
vote_count	Total number of votes in a day (including upvotes and downvotes)
user_count	Total number of individual users that have written posts or replies in a day
letter_count	the sum of letter counts in every post in a day
hot_degree	Compound index to evaluate the active degree of the day

Where the feature 'hot_degree' is calculated as:

$$hot\_degree = submission\_count * 7 + reply\_count * 1 + user\_count * 1 + vote\_count * 0.1 + letter\_count * 0.002$$

In summary, 18 features are created for each day.

5.2 Descriptive statistics and Predictive statistics

After extracting the features, several analyses are applied based on the features and the daily stock price change. The first is the descriptive statistics that contain the data visualization and Pearson correlation coefficient calculation. The second is predictive statistics. The features and prices are organized into a dynamic training set and testing set by using moving window measurement. After that, the machine learning models are trained and predict the prediction results. The prediction result performances are

evaluated by two aspects: the accuracy of direction prediction and an investment simulation based on prediction. The details of the training set and testing set construction are illustrated in later sections and the results of the analyses and evaluations are presented in Chapter 6. Figure 12 introduced the process of data analysis.

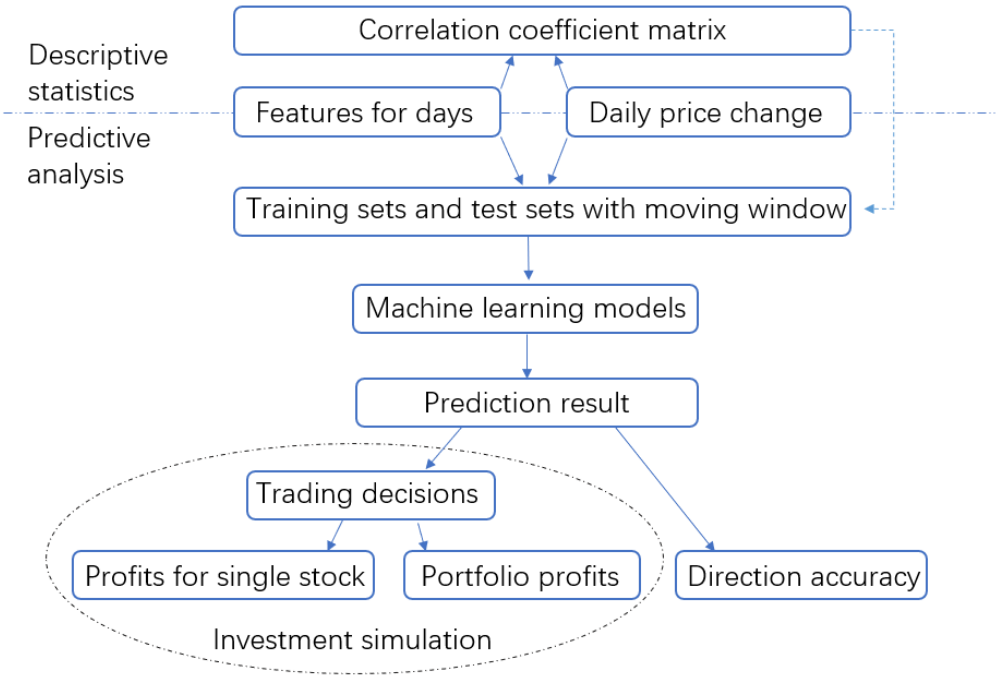


Figure 12 Process of Data analysis

The analysis starts with the features of each day and daily stock price change as input data. In the Descriptive Statistics part, which is on the top of the figure, a set of correlation coefficient matrix is presented to evaluate the linear correlation between the two input sets. The matrices are also used to help to construct the training set and determining the moving window length. In the predictive analysis, the data sets are used to create the training set and testing set with the 'moving window' technique. After that, Machine Learning models are built based on the training set, testing set, and several Machine Learning algorithms, which output the corresponding prediction results for each day. Last, the prediction results are evaluated from two aspects. The first one is the direction (grow, drop or steady) prediction accuracy. The second one is

the investment simulation, which is the bottom left circle in the figure. Firstly, the trading decisions are made based on the prediction results. After that, the trading simulations are made based on that trading decision under two different situations, trading with single stock and trading with Portfolio strategy. The profits based on two trading simulations are evaluated by comparing them with the profit of the Buy and Hold (B&H) baseline strategy.

5.2.1 Features and Label for sample

Feature selection is important for improving prediction performance. In addition to the quality of features, the number of features is also important since models need enough features to describe the data well while, on the other hand, the feature set should be as small and simple as possible to avoid overlearning. According to the linear correlation tests shown later in Chapter 6, it is reasonable to consider that the sentiment features of each day can influence the stock price changes in the next 4 days. Therefore, in models, when predicting the price change on a given day, all features in 4 days ahead will be used. The label of the given day and the features that in 4 days ahead consist of a sample that used in machine learning. The features and label building process for each sample is shown in

Figure 13 Features and label for sample below. For each available day (days with enough data to extract sentiment features and the stock market is open), for example, day 05.10 in the figure, the features of that day are consisted of all extracted features (based on Section 5.1.2) from 4 days ahead (01.10, 02.10, 03.10 are 04.10). So, each sample includes 72 features (4 days * 18 features per day) and one label which is generated using the close price of that day. All the samples are ordered based on the day which used to generate the label. For the convenience of the further discussion, we name each sample based on the date of label generated day (for instance, for the sample 1 in Figure 13, we call it: the sample of 05.10).

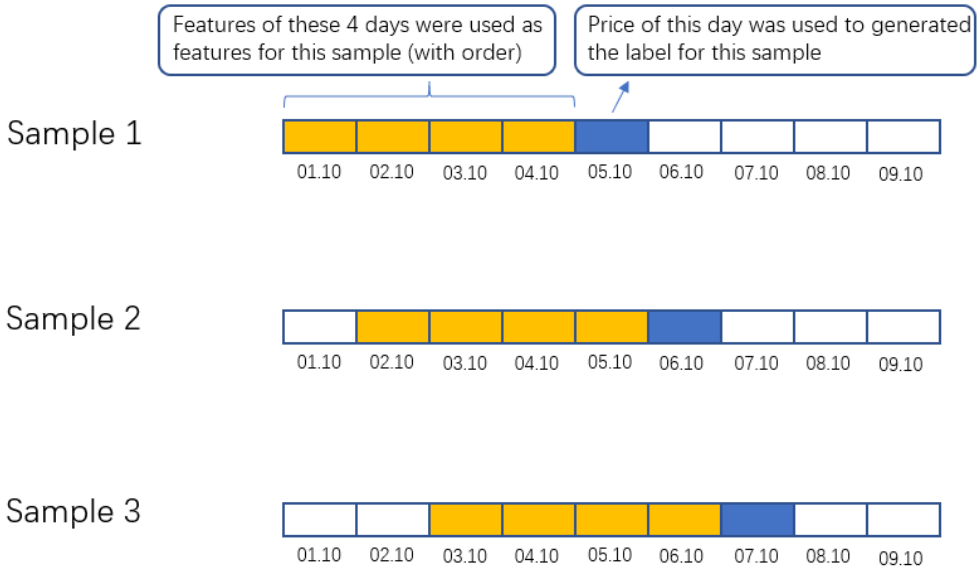


Figure 13 Features and label for sample

In addition, since activity-related features shown a strong linear correlation with stock prices in the Pearson correlation coefficient test (shown later in Section 6.1.2), an extra model using only the activity-related features in 4 days ahead is implemented.

In this study, the prediction performance measures are direction prediction accuracy (Section 6.2) and simulation investment profits (Section 6.3). For the direction prediction accuracy part, two different label setting measurements called 2-label and 3-label are used. In 2-label measurement, the day with positive price movement (where the closing price of the day is higher or equal to the previous day's closing price) are labeled as "grow", the day with negative price movement (where the closing price of the day is lower than the previous day's closing price) are labeled as "drop". In 3-label measurement, the samples are labeled as "grow", "drop" and "steady" where the closing price change are higher than +0.4% (increase from the previous day's closing price), lower than -0.4% (decrease from the previous day's closing price) or between +0.4% and -0.4%.

5.2.2 Training set and Testing set with moving window

In prediction using machine learning methods, the algorithms use past samples to predict future samples. However, as the data is a continuous time-series, the “past” and “future” is changing while the prediction process goes on. In order to solve this problem and keep the training samples up to date. A dynamic window sliding method is applied, where all data samples (samples built based on Section 5.2.1) from the beginning of 2015 (not the first day, but the first day that is available to build a sample) to the current window position were used for training and the samples of next 10 days were used for testing. The process is shown in Figure 14.

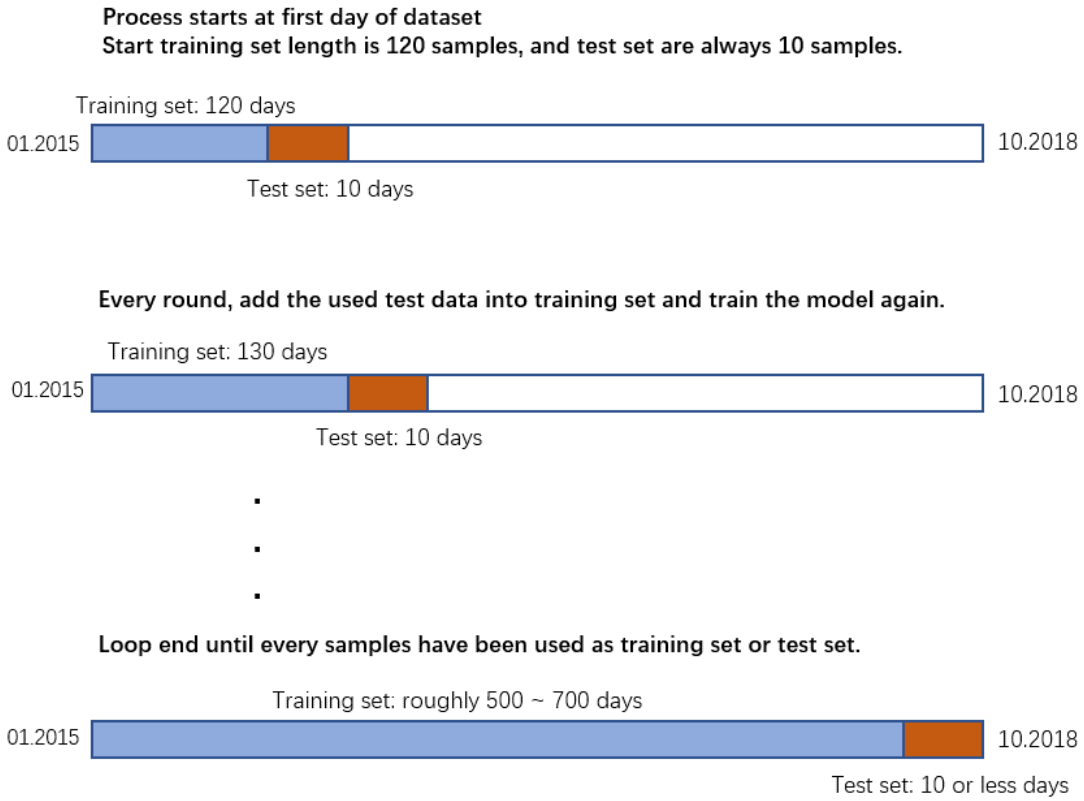


Figure 14 Moving window method

After data filtering and feature selection, all valid samples are ordered by their created date. The training window starts from the beginning of the time series with 120 samples, the testing window (length set as 10 samples) is adjacent following the training window. Machine learning models are trained and applied to predict using the current training

set and the testing set. After each round, the testing window moves forward for a given length (same length as the testing set) and the tested samples are added into the training set with their correct labels. The machine learning models are trained and be used to predict again using the new training set and testing set. The loop ended until every available sample have been used in the training set or testing set.

All the machine learning processes are implemented using the Scikit-learn machine learning package (Pedregosa et al., 2011) in Python programming language with the default setting.

In addition, for some machine learning models like Gradient Boosting Model (GBM) and Random Forest Classifier (RFC), the learning process involves randomness. In order to evaluate the models better and avoid the uncertainty from random seeds, the accuracy results, as well as the simulated profits, are the mean of 10 repeated experiments.

6 DATA ANALYSIS RESULTS

6.1 Descriptive statistical analysis results

Descriptive statistics provide simple but straightforward summaries or descriptions about the data with graphs and tables. The commonly used methods include the quantitative summarizing for single variance like means, variances, and density distribution graphs. However, there are also methods for multivariance descriptions like the multi-dimension scatter plots, the measures of dependence such as correlation measures and the conditional distributions.

In this chapter, the variable distributions are presented. Besides that, the Pearson correlation coefficient, as well as the correlation hypothesis tests, are implemented to analyze the linear correlations between the variables.

6.1.1 Feature distributions

Chapter 3 illustrated how sentiment analysis is implemented. In this section, the distributions about sentiment features (using the feature “vader_all_vote” as an example) and activity features (using the feature “hot_degree” as an example) are shown to provide overviews about the features and Reddit forum text data. The sentiment feature for each day in Figure 15 shows a distribution that visually similar to a normal distribution with stable mean and variance across the whole period. Figure 16 presents the activity of the forum for each day. It can be seen that the activity has a fluctuation in every 6 months. In addition, the forum is more active in March, September and less active in December, January, and July. Furthermore, the capacity of forum posts had a significant increase in 2018.

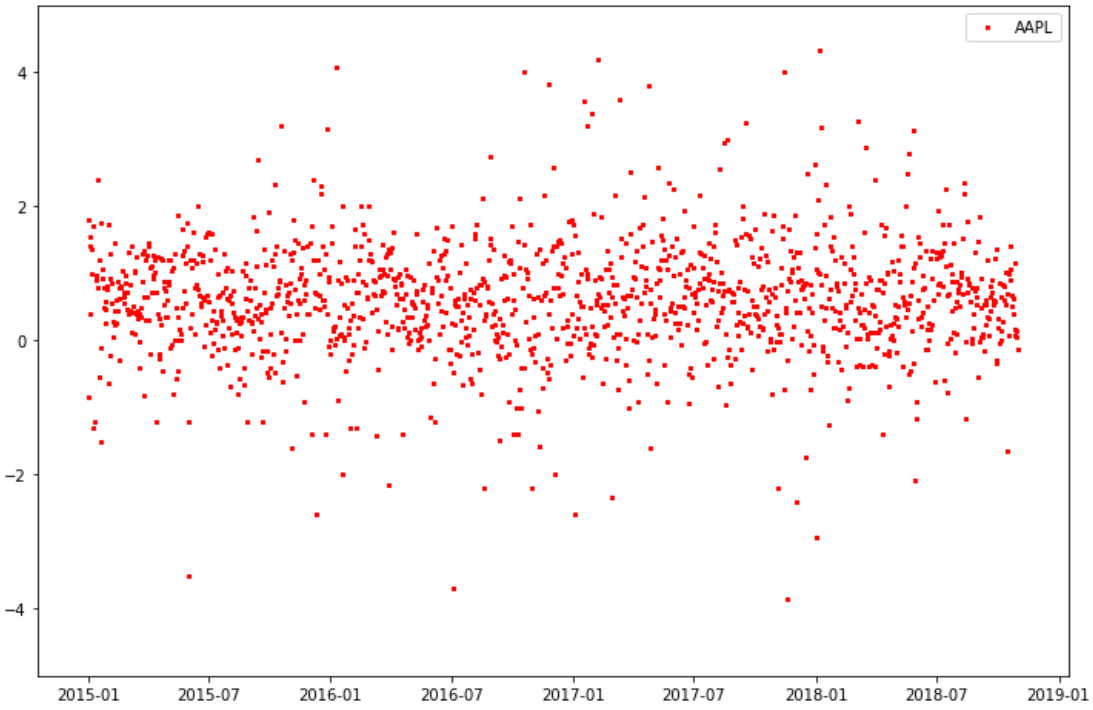


Figure 15 Sentiment score 'vader_all_time' distribution over time

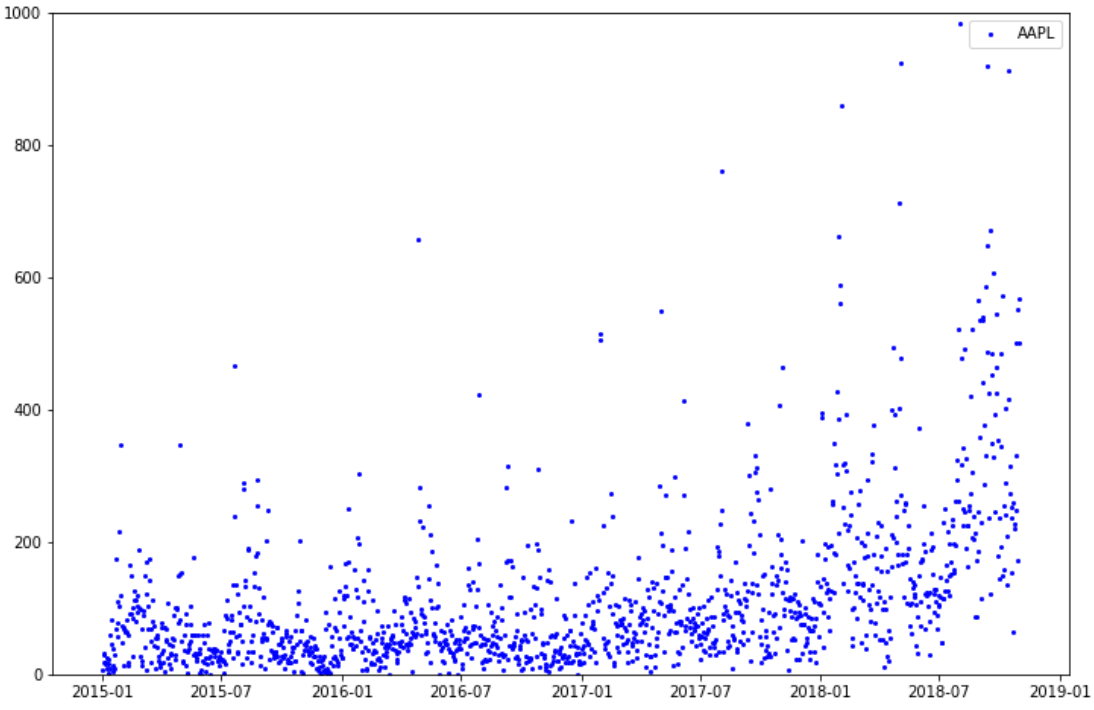


Figure 16 Feature "hot_degree" distribution over time

6.1.2 Correlation coefficient

In this section, the correlation coefficients between sentiment features and price changes in different day-lag are calculated.

Pearson correlation coefficient is a widely used statistical analysis method that measures the linear correlation between two values. An estimation of it is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (24)$$

Where n is the sample size, x_i and y_i are sample pairs of the two variables, and $\bar{x}$ and $\bar{y}$ are their sample averages.

In addition to the Pearson correlation coefficient, correlation hypothesis tests were also applied to evaluate the relationship between features. The statistical hypothesis test expressed whether the observations of samples are significant enough to confirm the hypothesis that the two variables are correlated. In the test, the p-value is used, which corresponds to the probability of the distribution of observation samples under the null hypothesis. The smaller the p-value is, the higher the probability it is to reject the null hypothesis of uncorrelated variables.

In this thesis, a t-test for nonzero correlation coefficient is used. It works as follows. When a t-test is used to test for correlation between two variables, the two variables are assumed to follow normal distributions and the null hypothesis is that the two variables are non-correlated. In this situation, the sample pairs should follow a bivariate normal distribution, where the sampling of the Pearson correlation coefficient between the variable pairs follows *t*-distribution with the degree of freedom $n-2$ (where *n* is the number of sample pairs). Specifically, the modified variable

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad (25)$$

follows the null case of *t*-distribution. Where *n-2* is the degree of freedom and *r* is the observed Pearson correlation coefficient.

The p-value of the test is expressed as the function below.

$$p - value = 2P(t > x|\mathrm{H}_0) = \int_{\mathrm{x}}^{+\infty} f(t)\mathrm{dt} \quad (26)$$

Where H_0 corresponds to the null hypothesis and

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{\mathrm{t}^2}{\nu})^{-\frac{\nu+1}{2}} \quad (27)$$

$$x = r\sqrt{\frac{n-2}{1-r_0}} \quad (28)$$

for an observed Pearson correlation coefficient value r_0, ν is the degree of freedom and

$$\Gamma(n) = (n-1)! \quad (29)$$

In the following tables, the significant levels thresholds were $p < 0.10$, $p < 0.05$ and $p < 0.01$, which were respectively marked as *, **, and *** after the Pearson correlation coefficient.

Table 9 is the correlation coefficient table for stock AAPL. As shown in the table, the correlation between feature “swn_mean” and the price change one day before (column ‘p=s-1’) is very significant correlated in the same direction. In addition, the activity-related features show a strong correlation with the price in one-day, four-day and five-day later so that a more active forum corresponds to a higher probability that the stock price will grow.

Table 10 shows the same thing in stock GOOG. Although the correlation between the “vader_var” and 3-days later stock price is significant at $p < 0.01$ level, it is hard to conclude any pattern from it. However, the activity-related features show a strong correlation with the price change in one-day, two-day and three-day later in reverse direction. Different from the pattern in stock AAPL, the more active the stock GOOG are discussed, the stock price is more likely to drop in 1,2 and 3 days.

Table 11 shows the stock MSFT correlations. From the column ‘p=s’, there are significant correlations between the stock prices and social media features on the same day, especially the “SentiWordNet” sentiment scores. The feature ‘submission

count' also shows that it is influenced by the price change on the same day and one day before.

As shown in future price columns, although there are 2 variable pairs significant in 0.05 level and 2 pairs in 0.10 level, from the overall view, it was still hard to summarize the relationship between forum features and prices as a behavior pattern or dependence. These significant results may come from the fault discovery rate in multiple hypothesis test situation, which is not reliable. Based on the linear correlation test, those features mentioned in the table had no linear correlation with stock MSFT price changes.

Table 9 Correlation coefficient table for stock AAPL

AAPL	p=s-1[7]	p=s[8]	p=s+1[9]	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	0.1074***	0.0032	0.0291	0.0319	0.0129	-0.0219	0.021
swn_var	0.0165	0.0389	-0.0345	0.0482	-0.0192	0.044	-0.0523
vader_mean	0.0207	0.0121	0.0307	-0.0105	-0.015	-0.0358	-0.0025
vader_var	0.0073	0.0043	-0.0573*	0.0103	0.017	0.0505	0.0446
swn_all_mean	-0.0031	0.0315	0.0316	0.0015	0.002	-0.0043	0.0191
swn_all_var	0.0171	0.0413	0.0166	0.0593*	-0.0027	0.0027	-0.0178
vader_all_mean	0.0256	0.0338	0.062*	-0.0251	-0.0153	-0.0235	-0.0094
vader_all_var	-0.0137	-0.0489	-0.0512	0.0439	0.0342	0.0254	0.08**
swn_vote_mean	0.0534	0.015	0.0278	0.0327	-0.0061	0.0013	0.032
vader_vote_mean	0.0056	0.0358	0.0012	0.0157	-0.0213	-0.0097	0.0237
swn_all_vote_mean	-0.0003	0.0144	0.032	0.0134	-0.0129	-0.006	0.0797**
vader_all_vote_mean	0.0371	0.0565*	0.0348	-0.0223	-0.0321	-0.0356	-0.0015
submission_count	-0.027	0.0282	0.0571*	-0.0021	0.0133	0.0656**	0.0414

[7] The notation 'p=s-1' means the sentiment features and activity-related features were calculated from forum in day X, while the stock price change data were collected from day X-1 (one day before day X).

[8] The notation 'p=s' means they were from same day.

[9] The notation 'p=s+1' means the sentiment features and activity-related features were calculated from forum in day X, while the stock price change data were collected from day X+1 (one day after day X). Where 'p=s+2' corresponds to same thing with '2 day after'. Same for 'p=s+3', 'p=s+4', 'p=s+5'.

reply_count	-0.0096	0.0401	0.0923***	0.043	0.0473	0.0717**	0.0475
vote_count	0.0094	0.0614*	0.1002***	0.014	-0.0018	0.1259***	0.0619*
user_count	-0.0085	0.0383	0.0904***	0.0392	0.0525	0.075**	0.0548*
words_count	-0.0267	0.0031	0.0402	0.0001	0.0126	0.0516	0.0745**
hot_degree	-0.0028	0.0492	0.092***	0.0199	0.013	0.1254***	0.0709**

Table 10 Correlation coefficient table for stock GOOG

GOOG	p=s-1	p=s	p=s+1	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	0.013	0.0368	0.0145	0.0068	0.0452	-0.0062	0.0121
swn_var	0.0273	0.0266	-0.0005	-0.0071	-0.0544	0.0026	0.0133
vader_mean	0.0157	0.0019	-0.0091	0.0492	0.0668*	0.0017	0.0149
vader_var	0.0156	0.0287	-0.0033	-0.0282	-0.0977***	-0.0341	-0.0601*
swn_all_mean	0.0686**	0.0217	0.0551*	-0.0135	0.0364	-0.0131	0.009
swn_all_var	-0.0026	0.0591*	0.0411	0.0431	0.0124	-0.07**	0.018
vader_all_mean	-0.0081	0.0041	0.0247	0.0381	0.0073	-0.0212	-0.0075
vader_all_var	-0.0014	0.0262	-0.0381	-0.0441	0.0094	-0.0175	0.0194
swn_vote_mean	-0.0318	0.0433	-0.0299	-0.0037	0.0135	0.0231	0.0397
vader_vote_mean	-0.0374	0.0041	-0.0052	0.0241	0.0261	0.0405	0.0419
swn_all_vote_mean	0.0337	0.0275	0.0611*	-0.013	0.0268	-0.0186	0.0336
vader_all_vote_mean	-0.0362	0.0203	0.0357	0.0184	0.0373	-0.0238	-0.0037
submission_count	-0.0334	0.0392	-0.0338	-0.0186	-0.1216***	-0.0456	0.0114
reply_count	-0.0059	-0.0331	-0.0396	-0.0422	-0.0735**	-0.0458	-0.0029
vote_count	-0.0232	-0.0539	-0.0923***	-0.0942***	-0.0765**	0.0094	0.01
user_count	-0.0152	-0.0271	-0.0386	-0.05	-0.0844***	-0.0461	-0.005
words_count	-0.0129	-0.0013	-0.072**	-0.0122	-0.0869***	-0.0541*	-0.0016
hot_degree	-0.0247	-0.0326	-0.0756**	-0.0646**	-0.104***	-0.0271	0.0066

Table 11 Correlation coefficient table for stock MSFT

MSFT	p=s-1	p=s	p=s+1	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	-0.0155	0.1156***	0.0211	-0.0388	0.0211	0.0581	-0.0364
swn_var	0.0281	-0.0007	-0.0354	-0.0437	0.0069	-0.0301	-0.0524
vader_mean	-0.0165	0.0275	-0.0135	0.0119	0.0222	0.0933**	-0.0317
vader_var	0.038	0.0057	0.0488	-0.0112	-0.0639*	-0.0327	-0.0536
swn_all_mean	-0.0359	0.0897***	-0.0465	-0.0299	0.0055	-0.0412	-0.0364
swn_all_var	0.0188	0.0293	0.0528	-0.0181	0.0288	-0.0085	-0.0067
vader_all_mean	-0.01	0.0402	-0.0171	-0.0362	0.0417	0.0481	-0.0248
vader_all_var	0.0427	-0.0242	-0.0071	0.0074	-0.0275	-0.0417	-0.0304
swn_vote_mean	-0.0038	0.0725**	0.0591	-0.001	0.0195	0.0439	-0.065*
vader_vote_mean	0.0102	0.0429	0.0353	-0.0233	0.0718*	0.0785**	-0.0554
swn_all_vote_mean	-0.032	0.0792**	-0.0127	-0.038	0.0404	0.002	-0.0086
vader_all_vote_mean	0.0222	0.0561	0.0159	-0.011	0.0476	0.0279	-0.0311
submission_count	-0.0662**	0.0677**	0.0441	-0.0337	0.0192	-0.0382	0.0095
reply_count	-0.0619*	-0.0291	0.0207	-0.011	-0.0061	-0.0361	-0.0366
vote_count	-0.017	0.036	0.0152	-0.0188	-0.0104	-0.0276	-0.0495
user_count	-0.0503	-0.0153	0.018	-0.0107	-0.0024	-0.033	-0.0268
words_count	-0.0298	0.0369	0.0048	0.0083	-0.0079	-0.0191	-0.0308
hot_degree	-0.0477	0.0195	0.0215	-0.0149	-0.0044	-0.036	-0.0376

Table 12 stands for stock NFLX. In this table, several sentiment-related features have a highly statistically significant relationship with future stock prices, no matter from comparing with the same features in other stocks' tests or from the significant level aspect.

The sentiment features, especially the sentiment scores from the SentiWordNet package have strong correlations with the 3-day future price change, higher sentiment scores might lead to the stock price growing in 3-day later. In addition, the activity-related features show an impact on 4-day future stock price change in a reverse direction.

From the correlations of stock TSLA in Table 13, although the sentiment-related features have no consistently significant correlations with historical or future price changes, the activity-related features are statistically significantly correlated with the 3-day future price change in the same direction.

Table 14 presents the correlation coefficients between stock AMZN features and price changes. The sentiment-related features “swn_all_mean”, “swn_all_vote_mean” and “vader_all_vote_mean” have significant correlations with the 4-day future price change. However, it is interesting that all significant variable pairs with the 4-day future price change have minus correlation coefficient value which means that the public optimistic may lead to stock price decreases. Besides that, the activity-related features revealed a strong correlation in 3-day ahead in the opposite direction.

Table 12 Correlation coefficient table for stock NFLX

NFLX	p=s-1	p=s	p=s+1	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	-0.023	0.0161	-0.0043	-0.0172	0.1185***	0.014	-0.0442
swn_var	0.015	0.0356	0.0134	0.0829**	-0.0609*	-0.0729**	-0.0036
vader_mean	-0.0011	0.0021	0.0194	-0.0443	0.0668*	0.019	0.0466
vader_var	-0.001	0.0414	-0.0391	0.0131	-0.0449	-0.0154	-0.0533
swn_all_mean	0.0166	0.021	0.0064	-0.0425	0.0678*	0.0246	0.0125
swn_all_var	0.0407	0.0463	-0.0076	0.065*	-0.0712**	-0.0417	-0.0085
vader_all_mean	0.0288	0.0107	-0.0346	-0.0159	-0.0099	-0.0458	0.063*
vader_all_var	0.0098	0.0197	0.0175	-0.018	0.0124	0.0141	-0.0148

swn_vote_mean	-0.0086	-0.0047	0.0147	-0.0263	0.106***	-0.0209	-0.0179
vader_vote_mean	0.0205	0.0279	0.0272	-0.0152	0.0494	0.0381	0.0168
swn_all_vote_mean	-0.0022	0.0049	0.0047	-0.0025	0.0937***	0.0411	0.0405
vader_all_vote_mean	0.0117	0.0141	0.0043	-0.0133	-0.0072	-0.0069	0.0509
submission_count	0.0295	0.0152	-0.0109	-0.0356	-0.0206	-0.0551*	-0.0215
reply_count	0.0322	0.0082	-0.0479	-0.0585*	-0.0274	-0.0902***	-0.0447
vote_count	0.0285	0.0081	-0.0267	-0.0434	-0.013	-0.0251	-0.0392
user_count	0.0332	0.0122	-0.0404	-0.0529	-0.0257	-0.082**	-0.0362
words_count	0.0405	0.0685**	-0.0493	-0.0349	-0.0102	-0.0545*	-0.0612*
hot_degree	0.036	0.0193	-0.0349	-0.0469	-0.0202	-0.0649**	-0.048

Table 13 Correlation coefficient table for stock TSLA

TSLA	p=s-1	p=s	p=s+1	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	-0.0009	-0.0124	-0.0778**	0.0297	-0.0109	0.0303	0.0036
swn_var	0.0008	0.0189	0.0177	-0.0664*	0.0079	0.0287	-0.0236
vader_mean	0.0607*	0.0101	-0.0316	0.0305	-0.0504	-0.0101	0.026
vader_var	-0.0389	-0.0173	-0.0037	-0.0147	-0.0037	0.0309	-0.0026
swn_all_mean	-0.0357	-0.0114	-0.0445	0.0006	-0.0344	-0.0157	-0.0028
swn_all_var	0.0059	-0.018	0.0156	-0.0207	-0.0197	0	0.0192
vader_all_mean	0.0159	0.0023	-0.0208	0.0117	0.0032	0.0098	0.0439
vader_all_var	-0.0149	-0.0408	0.0291	-0.0176	0.0173	-0.0007	-0.0243
swn_vote_mean	0.0477	0.0166	-0.0447	-0.0021	0.0414	-0.0067	-0.0072
vader_vote_mean	0.073**	0.0307	-0.0482	-0.0046	0.0352	-0.0123	-0.013
swn_all_vote_mean	0.0147	0.0208	-0.0291	0.0073	0.0253	-0.024	-0.0273
vader_all_vote_mean	0.0416	0.05	-0.0185	0.0047	0.0278	0.0241	0.0264
submission_count	0.0277	0.0384	0.0301	0.011	0.0809**	0.0035	0.0457

reply_count	0.0538	-0.0037	0.0478	0.0186	0.0966***	0.0158	0.0595*
vote_count	0.0002	-0.0041	0.0017	-0.0175	0.0295	-0.0142	0.0053
user_count	0.0445	-0.0092	0.0389	0.013	0.0898***	0.0115	0.0488
words_count	0.0468	-0.0272	0.0365	-0.0041	0.0684**	0.0206	0.0241
hot_degree	0.0232	0.0009	0.0217	-0.0044	0.062*	-0.0007	0.0293

Table 14 Correlation coefficient table for stock AMZN

AMZN	p=s-1	p=s	p=s+1	p=s+2	p=s+3	p=s+4	p=s+5
swn_mean	-0.0291	0.0078	0.0232	0.0024	0.042	-0.0168	0.0634*
swn_var	-0.0147	0	0.0324	-0.016	-0.0019	-0.0533	-0.0577*
vader_mean	0.0415	0.0447	-0.0762**	-0.0192	0.0531	-0.0198	-0.035
vader_var	-0.037	0.0198	0.0081	-0.0223	-0.0302	-0.0051	-0.0086
swn_all_mean	0.0072	0.0076	-0.0019	0.0151	0.0152	-0.09***	0.0167
swn_all_var	-0.0295	-0.0086	0.0156	-0.0463	-0.0313	-0.0584*	-0.0142
vader_all_mean	0.0611*	0.0615*	0.0641*	-0.0043	0.0553	-0.0551	-0.0117
vader_all_var	-0.0357	0.0104	-0.0301	-0.0154	-0.0368	-0.0341	0.005
swn_vote_mean	-0.0002	-0.0288	-0.0077	0.0204	0.0471	-0.0297	0.0539
vader_vote_mean	-0.0153	0.0523	-0.0379	-0.0265	0.0135	0.0293	-0.0347
swn_all_vote_mean	0.0473	-0.01	0.0322	0.0221	-0.013	-0.1047***	0.0153
vader_all_vote_mean	0.1022***	0.0483	0.0498	0.0332	0.0257	-0.0734**	-0.0002
submission_count	-0.0177	0.0062	-0.0101	-0.067**	-0.0952***	-0.0441	-0.006
reply_count	-0.0175	-0.0602*	-0.0548*	-0.0674**	-0.1149***	-0.0684**	-0.0372
vote_count	-0.0273	-0.046	-0.035	-0.0509	-0.0704**	0.0019	-0.019
user_count	-0.0207	-0.0469	-0.0441	-0.0716**	-0.1054***	-0.0642*	-0.0329
words_count	-0.0039	-0.03	-0.0558*	-0.0333	-0.1008***	-0.0634*	-0.0394
hot_degree	-0.0227	-0.0433	-0.0429	-0.0639*	-0.1043***	-0.0403	-0.0277

Generally speaking, in most of the stocks except MSFT, there are statistically significant correlations between activity-related features and price changes in roughly 3 or 4 days ahead which reflects the potential of prediction power. Furthermore, for those sentiment-related features, only those in one stock (NFLX) have significant enough correlations with 3-day ahead price change. However, for other stocks, although there are some feature pairs that have small p-values, the significant correlations between sentiment scores and price change are not consistently located to be considered as a regular pattern. In addition, since the tests above are all multiple hypothesis tests, the fault discovery rate measurement might help to adjust the threshold to make more tests significant.

Although all the test data and users are from Reddit online forum, different stocks show different characteristics that the correlation direction between the activity-related features and price changes is not the same. The activity of discussion about AAPL and TSLA in the forum may lead the stock price increase while the activity of GOOG, NFLX and AMZN posts might cause the price decrease.

6.2 Predictive analysis results

Predictive analysis is a broad term of analytical methods and techniques using historical samples to predict future or unknown outcomes and behaviors (Nyce & Cpcu, 2007). The key idea of the predictive models is that the historical data and future data should have similar behavior. Being represented by the machine learning algorithms and regression methods, predictive models capture the relationships or patterns from historical data and use them in estimating the new input data.

The predictive analysis has been applied in many fields like marketing, insurance, and financial services. Companies or organizations usually directly benefit from the prediction results. Besides that, predictive analysis can also be used to validate the relationships between variables based on the performance of prediction on the known data. Good performance (for example, the high accuracy of prediction) means that the

model can extract the pattern clearly. On the other side, bad performance represents that patterns cannot be found or not exist.

Many methods can be used in predictive analysis, for instance, linear regression models and time series models. In this study, machine learning algorithms are used. Non-linear predictors learned by machine learning may be able to reveal hidden non-linear correlation or pattern that cannot be found in traditional linear methods. In this thesis, the accuracies of machine learning models on different stocks are examined. In addition, a simulated algorithm trading strategy is designed based on predictions, the overall profits of it are also investigated as part of the performance evaluation method.

In this chapter, the methods, models, and results of prediction and simulated algorithm trading are shown and explained.

6.2.1 Direction prediction accuracy

Direction prediction accuracy is one of the performance measures. In this section, the benchmark accuracy of the 2-label test is set as the proportion of the majority label. It means that a model that always predicts the overall majority label will be correct that proportion of samples. In addition, the accuracies of prediction are simply the percentage of correct predictions. However, in the 3-label test, the benchmark accuracies and prediction accuracies are based on a confusion matrix (Table 15).

Table 15 Confusion matrix for 3-label prediction

	True drop (-1)	True steady (0)	True grow (1)
Predicted drop(-1)	A(-1, -1)	A(0, -1)	A(1, -1)
Predicted steady(0)	A(-1, 0)	A(0, 0)	A(1, 0)
Predicted grow(1)	A(-1, 1)	A(0, 1)	A(1, 1)

The A(-1, -1) means the numbers of samples that belong to each category.

The benchmark accuracy is defined as:

$$\text{b\_acc} = \frac{\max[\text{True drop}, \text{True grow}]}{\text{True drop} + \text{True grow}} \tag{30}$$

Where

$$\text{True drop} = \text{A}(-1,1) + \text{A}(-1,0) + \text{A}(-1,1) \tag{31}$$

$$\text{True grow} = \text{A}(1,-1) + \text{A}(1,0) + \text{A}(1,1) \tag{32}$$

And the prediction accuracies are defined as:

$$\text{p\_acc} = \frac{\text{A}(-1,-1)+\text{A}(1,1)}{\text{A}(-1,-1)+\text{A}(-1,1)+\text{A}(1,-1)+\text{A}(1,1)} \tag{33}$$

In Table 16, Table 17, Table 18 and Table 19. The direction prediction accuracies of all 4 tests are presented. From the model aspect, the model LR, DT, RFC and GBM have generally better performance than others by comparing them with their benchmark accuracy, however, none of them have consistent leading performance in every stock prediction. From the stock aspect, the stock NFLX and AMZN are hard to predict that none or little of the machine learning model has accuracy higher than benchmarks. On the other side, the TSLA, GOOG, and AAPL are relatively more predictable that there exist several attempts that get results that better than benchmarks.

Although the activity-related features show strong linear correlations with future price changes, the direction prediction accuracies of them do not show evident leading in prediction power than using all features. Furthermore, comparing with 2-labels test, the 3-label tests perform better in direction prediction while they make the prediction more unstable that the best model achieves higher accuracy (e.g. AAPL got 0.5769 in Table 19) while the worst model becomes lower (NFLX got 0.4375 in Table 18).

Table 16 Direction accuracy / 2 labels / all features[10]

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average[11]
KNN	0.4992	0.5332	0.4688	0.4598	0.4936	0.5065	0.4935
SVM(linear)	0.5218	0.4735	0.5278	0.5123	0.4990	0.5178	0.5087
SVM(poly)	0.4978	0.5059	0.4983	0.4537	0.5263	0.4934	0.4959
LR	0.5049	0.4974	0.5540	0.5092	0.5027	0.5215	0.5150
DT	0.5168	0.5241	0.4822	0.4620	0.5139	0.5249	0.5040
RFC	0.4918	0.5279	0.5157	0.5055	0.4851	0.5296	0.5093
GBM	0.5111	0.5112	0.5032	0.5	0.5094	0.5452	0.5134
Average[12]	0.5062	0.5105	0.5071	0.4861	0.5043	0.5198	0.5057[13]
benchmark	0.5302	0.5451	0.5381	0.5123	0.5027	0.5703	

Table 17 Direction accuracy / 2 labels / only active features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
KNN	0.4852	0.5049	0.4827	0.4876	0.5073	0.5098	0.4963
SVM(linear)	0.4901	0.5098	0.5160	0.5	0.5221	0.5049	0.5072
SVM(poly)	0.4913	0.5270	0.4963	0.4987	0.5147	0.4704	0.4997
LR	0.5221	0.5073	0.4926	0.4802	0.4729	0.5307	0.5010
DT	0.5317	0.5123	0.5013	0.5222	0.4948	0.5352	0.5163
RFC	0.5083	0.5321	0.5220	0.5118	0.4938	0.5096	0.5129
GBM	0.5428	0.55	0.4887	0.4831	0.5128	0.5174	0.5158
Average	0.5102	0.5205	0.4999	0.4977	0.5026	0.5111	0.5070
benchmark	0.5302	0.5451	0.5381	0.5123	0.5027	0.5703	

[10] GBM = Gradient Boosting Model. SVM(linear) = Support Vector Machine with 'linear' kernel. SVM(poly) = Support Vector Machine with 'polynomial' kernel. LR = Logistic Regression. DT = Decision Tree Classifier. RFC = Random Forest Classifier. KNN = k-nearest Neighbor Classifier.

[11] The Average column at the right of the table is the average accuracy with respect to the different stocks with same model.

[12] The Average lane at the bottom of the table is the average accuracy with respect to the same stocks with different models.

[13] This cell means the average of all 42 tests with 6 stocks and 7 models. It reflects the quality of this feature set and label set.

Table 18 Direction accuracy / 3 labels / all features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
KNN	0.4823	0.5120	0.4420	0.4375	0.4988	0.4586	0.4719
SVM(linear)	0.4713	0.4867	0.5064	0.4644	0.4763	0.4982	0.4839
SVM(poly)	0.4936	0.5163	0.5602	0.4930	0.5259	0.4473	0.5061
LR	0.5762	0.5393	0.5205	0.4876	0.5215	0.5468	0.5320
DT	0.5252	0.5417	0.5626	0.5028	0.4983	0.5071	0.5230
RFC	0.5106	0.5224	0.4607	0.5038	0.5119	0.5325	0.5070
GBM	0.5089	0.5490	0.4643	0.4972	0.5298	0.5046	0.5090
Average	0.5097	0.5239	0.5024	0.4838	0.5089	0.4993	0.5047
benchmark	0.5498	0.5784	0.5566	0.5100	0.5173	0.5851	

Table 19 Direction accuracy / 3 labels / only active features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
KNN	0.4836	0.5147	0.4942	0.4803	0.5226	0.4955	0.4985
SVM(linear)	0.4792	0.4738	0.4822	0.5211	0.4964	0.5098	0.4938
SVM(poly)	0.5149	0.55	0.5173	0.5293	0.4937	0.5	0.5175
LR	0.5472	0.5314	0.5043	0.4801	0.4837	0.5277	0.5124
DT	0.5317	0.5044	0.5509	0.5072	0.4940	0.4877	0.5127
RFC	0.5506	0.5454	0.5307	0.4967	0.4983	0.5320	0.5256
GBM	0.5769	0.5196	0.5013	0.5006	0.4916	0.5169	0.5178
Average	0.5263	0.5199	0.5116	0.5022	0.4972	0.5099	0.5112
benchmark	0.5498	0.5784	0.5566	0.5100	0.5173	0.5851	

6.3 Simulated trading experiment

The simulated trading experiment is a measure to evaluate the predictability and model performance in the stock prediction area. In this section, an experiment based on the direction prediction results are implemented. In order to evaluate the prediction and trading strategy performance, the measures “total profit” and “daily profit efficiency” under each model are calculated and compared with a benchmark strategy to examine the predictability.

6.3.1 Simulated trading for single stock

In this section, the simulated trading strategies and benchmark strategies are designed for a single stock. For the simulation of a given stock, both simulated trading strategy and benchmark strategy starts with an assumed account with the same amount of money. The trading performance is evaluated by the overall profits at the end of the period.

In simulated trading strategies, each trade decision is made at the market close time of every day. The decisions are about whether to own the stock till the next day's market close time. If yes, buy stocks with all money when no stocks in hand or hold the owned stock for one more day. If no, sell all owned stock or do nothing if the account has no stocks. The trading strategy is described as the table below.

	Predict result	If account own stock	If account own no stock
2-label	grow	Do nothing	Buy stocks with all money
	drop	Sell all stocks	Do nothing
3-labels	grow	Do nothing	Buy stocks with all money
	steady	Do nothing	Do nothing
	drop	Sell all stocks	Do nothing
benchmark	-	Buy stocks with all money at the beginning and hold them forever.	

Table 20 Trading strategy

The benchmark strategy is simply buying the stocks with all money at the beginning of the period and holding it until the end of the period. This benchmark strategy is commonly used in stock prediction research called Buy and Hold Strategy (B&H).

The first evaluation measure is "total profit". The total profit is the profit ratio (final assets value divided by original value) after applying the trading strategy for a time

period (In this study, the time period is roughly 3 years to 1.5 years based on the data availability).

Table 21, Table 22, Table 23 and Table 24 shows the total profits in simulations where the former number in each cell is the overall profit ratio while the latter percentage is the proportion of that ratio to the corresponding ratio of Buy and Hold (B&H) benchmark.

Table 21 Total profit based on the prediction with 2 labels / all features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN
KNN	1.492/79.3%	1.311/108.5%	1.095/80.0%	0.885/51.9%	1.136/82.6%	1.834/87.2%
SVM(linear)	2.089/111.0%	0.985/81.5%	1.354/98.9%	1.34/78.7%	1.101/80.1%	1.518/72.2%
SVM(poly)	1.534/81.6%	1.019/84.4%	1.16/84.7%	0.949/55.7%	1.272/92.5%	1.652/78.6%
LR	1.869/99.4%	1.16/96.0%	1.355/99.0%	1.649/96.8%	0.936/68.1%	1.651/78.5%
DT	1.571/83.5%	1.287/106.5%	1.035/75.6%	1.203/70.6%	2.023/147.1%	1.665/79.2%
RFC	1.643/87.3%	1.281/106.0%	1.109/81.0%	1.317/77.3%	0.866/62.9%	1.991/94.7%
GBM	1.684/89.5%	1.211/100.2%	1.138/83.1%	1.639/96.2%	1.193/86.7%	2.122/100.9%
benchmark	1.881/100.0%	1.208/100.0%	1.369/100.0%	1.704/100.0%	1.375/100.0%	2.103/100.0%

Table 22 Total profit based on the prediction with 2 labels / only active features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN
KNN	1.278/76.0%	1.295/68.4%	1.451/61.9%	1.369/48.2%	1.23/93.0%	2.438/76.3%
SVM(linear)	1.07/63.6%	1.522/80.4%	1.755/74.8%	1.591/56.0%	2.578/194.9%	1.579/49.4%
SVM(poly)	1.173/69.7%	1.645/86.9%	1.3/55.4%	1.878/66.1%	1.355/102.5%	1.636/51.2%
LR	1.611/95.8%	1.433/75.7%	1.372/58.5%	1.059/37.2%	0.654/49.5%	3.296/103.1%
DT	1.644/97.7%	1.702/89.9%	1.865/79.5%	1.401/49.3%	1.085/82.1%	2.737/85.6%
RFC	1.873/111.4%	1.968/103.9%	1.88/80.2%	2.213/77.9%	1.447/109.4%	2.463/77.1%
GBM	2.153/128.0%	2.111/111.4%	0.925/39.4%	0.972/34.2%	1.68/127.0%	2.575/80.6%
benchmark	1.682/100.0%	1.894/100.0%	2.345/100.0%	2.843/100.0%	1.323/100.0%	3.196/100.0%

Table 23 Total profit based on the prediction with 3 labels / all features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN
KNN	1.117/59.4%	1.256/103.9%	0.938/68.5%	0.742/43.5%	1.114/81.0%	1.241/59.0%
SVM(linear)	1.227/65.2%	1.122/92.9%	1.144/83.5%	0.998/58.6%	1.002/72.8%	1.455/69.2%
SVM(poly)	1.257/66.8%	1.132/93.7%	1.283/93.7%	1.281/75.2%	1.163/84.6%	1.099/52.3%
LR	2.222/118.2%	1.362/112.8%	1.085/79.3%	1.281/75.2%	1.444/105.0%	1.741/82.8%
DT	1.512/80.4%	1.22/101.0%	1.162/84.9%	1.571/92.2%	0.908/66.0%	1.484/70.6%
RFC	1.441/76.6%	1.116/92.4%	0.931/68.0%	1.114/65.4%	1.314/95.5%	1.409/67.0%
GBM	1.5/79.7%	1.328/109.9%	1.007/73.6%	0.952/55.9%	1.232/89.6%	1.145/54.5%
benchmark	1.881/100.0%	1.208/100.0%	1.369/100.0%	1.704/100.0%	1.375/100.0%	2.103/100.0%

Table 24 Total profit based on the prediction with 3 labels / only active features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN
KNN	1.122/66.7%	1.333/70.4%	0.987/42.1%	1.254/44.1%	1.321/99.8%	1.917/60.0%
SVM(linear)	1.162/69.1%	1.186/62.6%	1.054/44.9%	2.303/81.0%	0.771/58.3%	1.232/38.5%
SVM(poly)	1.331/79.2%	0.93/49.1%	1.543/65.8%	2.204/77.5%	1.76/133.1%	1.842/57.6%
LR	1.61/95.7%	1.488/78.6%	1.524/65.0%	1.199/42.2%	0.98/74.1%	2.047/64.0%
DT	1.864/110.9%	1.155/61.0%	1.506/64.2%	2.563/90.1%	1.322/99.9%	1.182/37.0%
RFC	1.401/83.3%	1.64/86.6%	1.428/60.9%	1.755/61.7%	1.452/109.7%	2.003/62.7%
GBM	2.139/127.2%	1.655/87.4%	1.417/60.4%	2.048/72.1%	1.138/86.0%	2.026/63.4%
benchmark	1.682/100.0%	1.894/100.0%	2.345/100.0%	2.843/100.0%	1.323/100.0%	3.196/100.0%

Besides the profit tables, the figures of the simulated investments are also presented to compare the predictability of each stock and the predict performance for 4 tests with different features and label sets. As shown in Figure 17 to Figure 22. The curve for each test representing the best-performed machine learning model prediction results under that features and labels setting. In addition. the benchmark (B&H) profits are also shown in the figure.

It can be noticed that the curves of “all features” tests start roughly from 2016. This is because the forum data in early years are not intensive enough to be built as samples, in other words, early years data cannot be used in this situation (for instance, if there

is no post or reply for a stock in one day, then the sentiment scores for that day are not available). In addition, the activity-related features tests require less data than “all features” test (0 posts in one day can be used to calculate the activity-related features but not the sentiment scores) which makes them start earlier. The trading strategies of the later test are set as B&H when it is not available to predict.

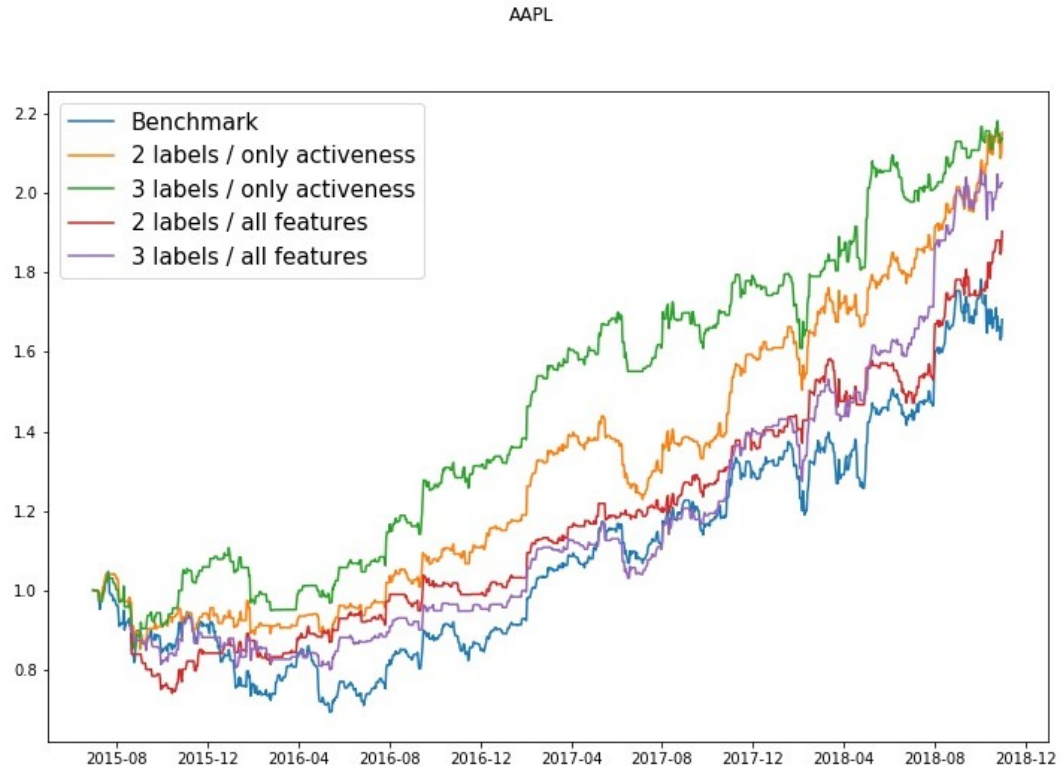


Figure 17 Profit plot for stock AAPL

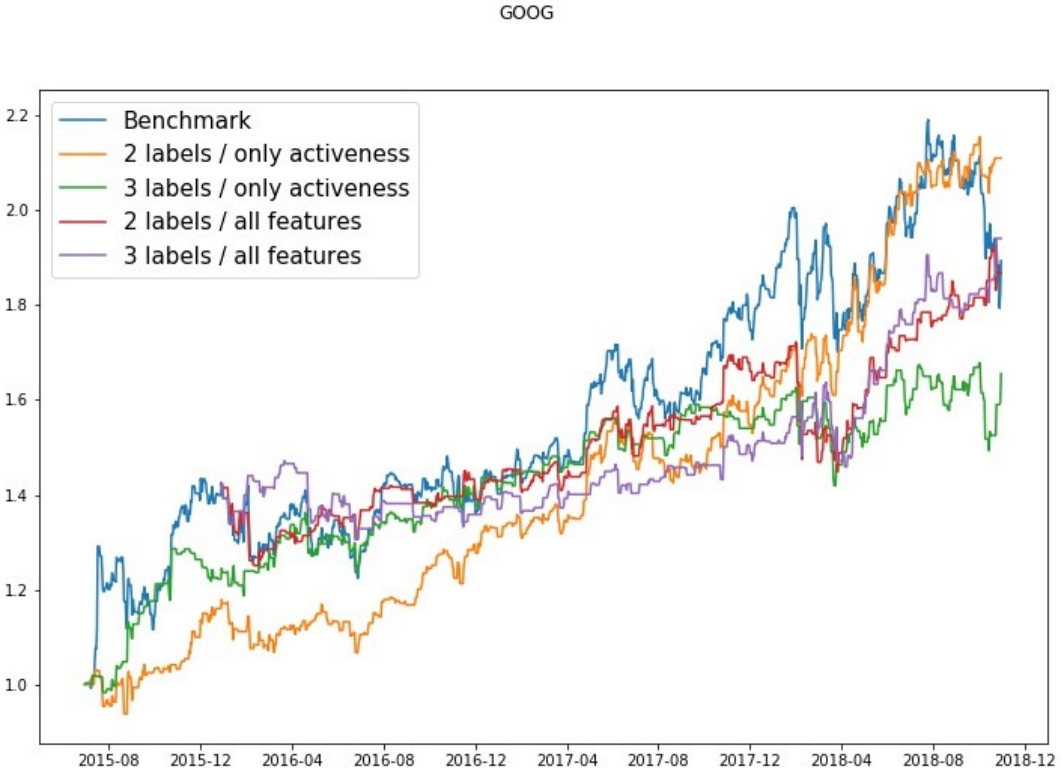


Figure 18 Profit plot for stock GOOG

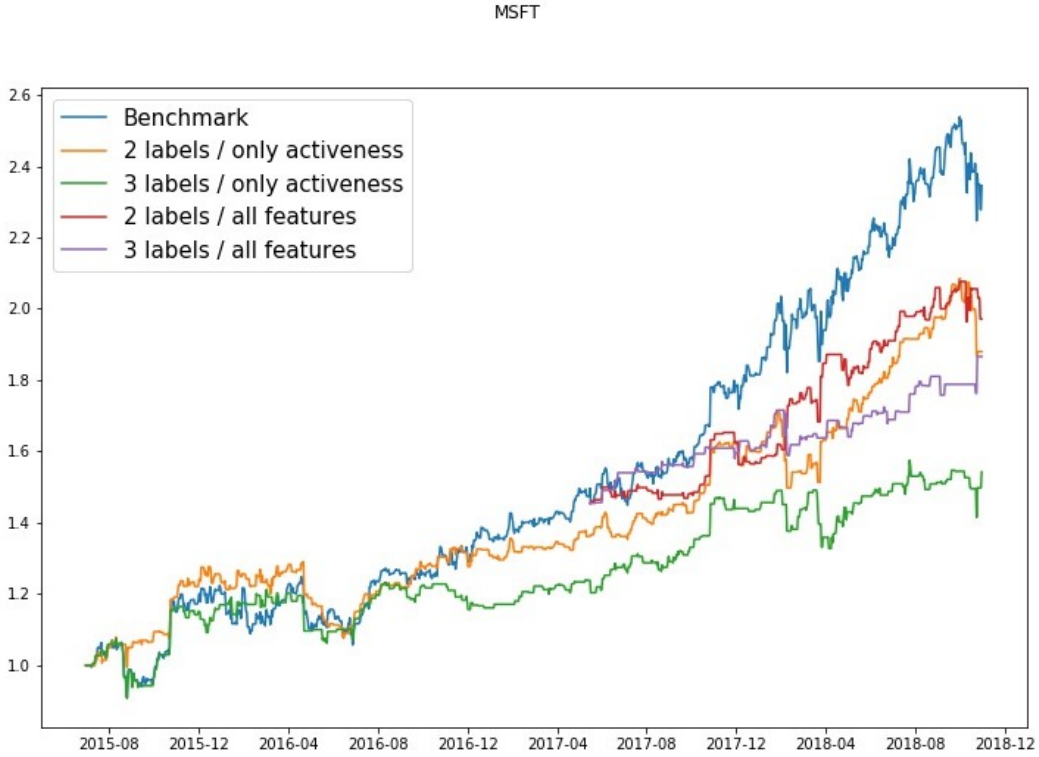


Figure 19 Profit plot for stock MSFT

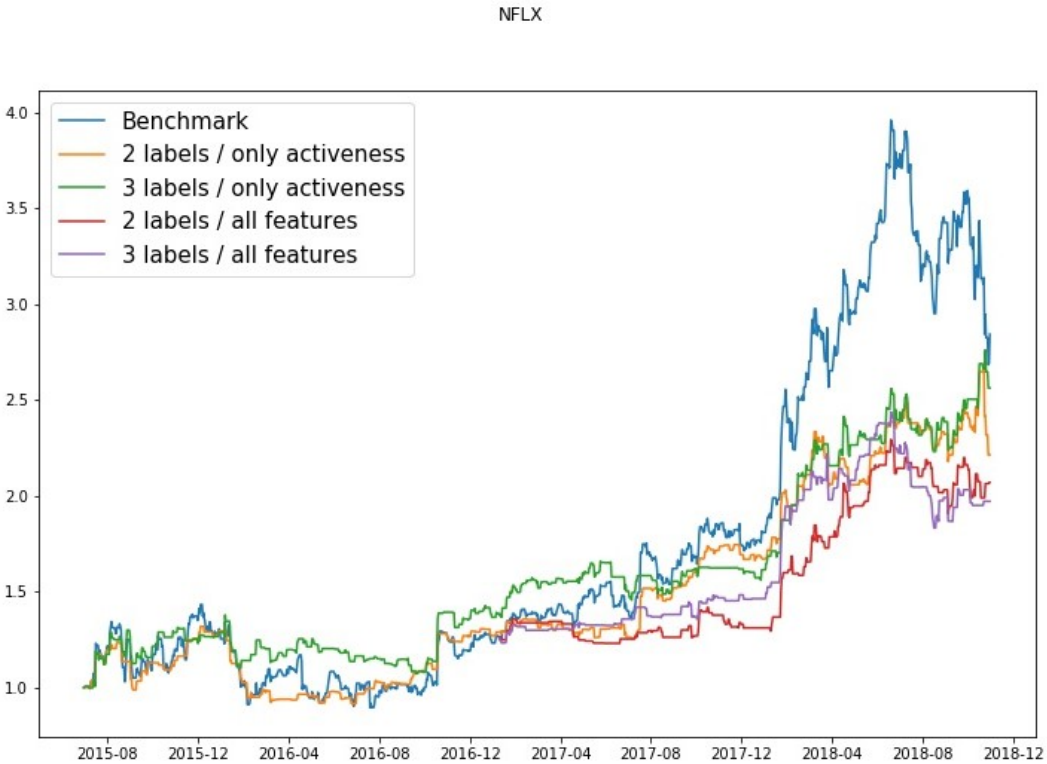


Figure 20 Profit plot for stock NFLX

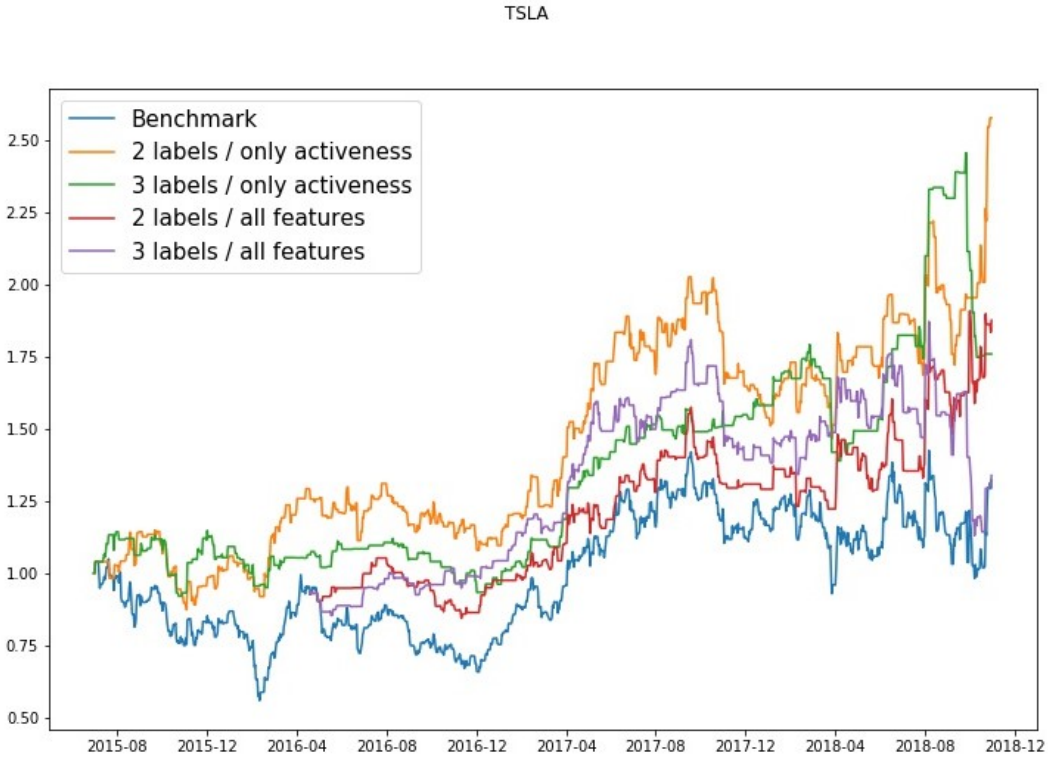


Figure 21 Profit plot for stock TSLA

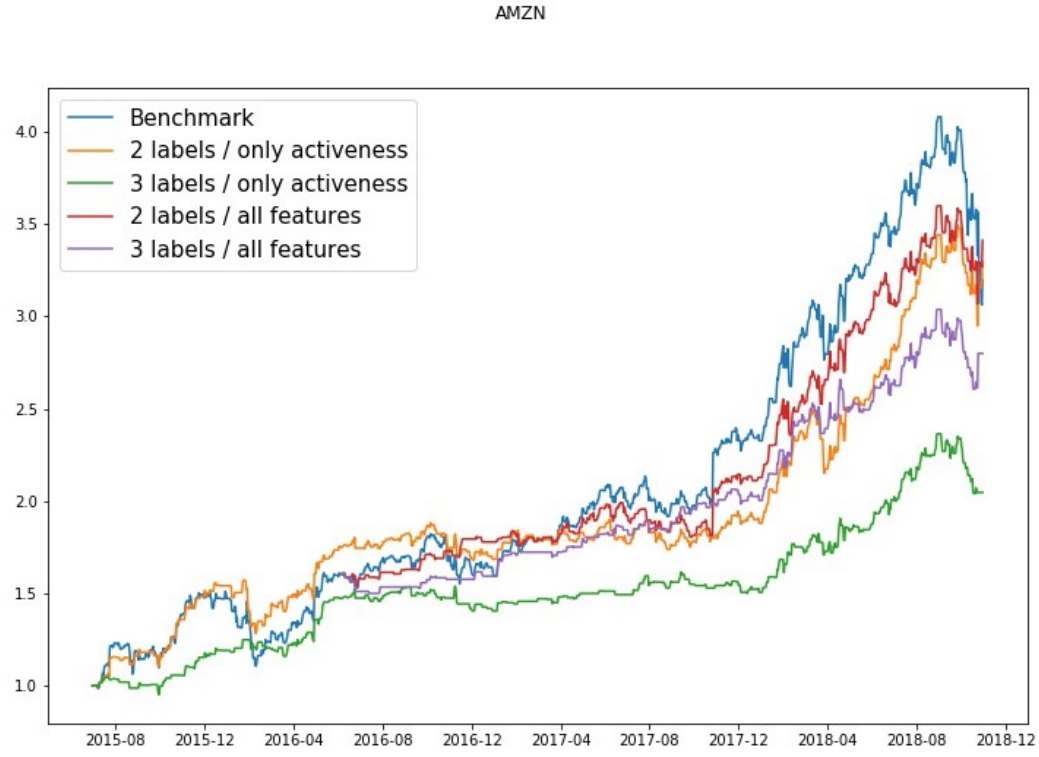


Figure 22 Profit plot for stock AMZN

As shown in the figures above. The stocks AAPL and TSLA are the most predictable stocks that there exist predictions get profits better than B&H strategy while the B&H

strategies are leading in all the other stocks. However, the different feature selection methods do not bring a strong difference in total profits.

The second evaluation factor is “daily profit efficiency”. In this thesis, it means the profit ratio if the money is invested in a stock for one day. It is calculated by total profit divided by numbers of day that strategy choose to own the stock instead of cash. For instance, benchmark strategy owns the stock all the time in the period while algorithms trading strategies only hold the stock while the predictions are ‘grow’ for the next day.

The idea of ‘profit efficiency’ has been used by many investors when they have multiple investment opportunities (for example, different stock in stock markets, Foreign exchange markets, Futures contract markets or even interest of deposit in banks). Investing assets into one market means that losing the opportunity to invest that assets in other places, so the investors prefer to invest the market with the highest profit efficiency every day.

This factor is applied in this thesis because the majority of stocks are up trending, the algorithm trading strategies get heavier punishment for not catching the growing day and less reward for avoid dropping day. Or in other words, every time the strategies choose not to trade on someday means losing profit in expectation which is ‘unfair’ to compare the total profit with B&H under this situation. The factor “daily profit efficiency” focuses more on the efficiency and assets utilization ratio.

Table 25, Table 26, Table 27 and Table 28 show the “daily profit efficiency” based on the predictions under different machine learning models and features selections. The percentages mean how many ratios can investors earn per day if invest one-unit of money in that stock on days when the algorithm decides to trade. As shown in tables. All the predictions perform well comparing with benchmark (B&H) strategies. Different from what “total profit” tables shown, the different feature selections show a significant difference in performance. The tests with 3-labels perform better than 2-labels tests while the predictions with “all features” provide more efficient profits than the predictions with “only activeness features”. In the machine learning model aspect, the model KNN, SVC-poly, and RFC perform better compared with other models.

Table 25 Daily profit efficiency with 2 labels / all features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average[14]
GBM	0.3890%	0.2896%	0.6394%	0.9262%	0.3487%	0.5498%	0.5238%
SVC	0.4694%	0.2773%	0.7320%	0.7447%	0.3507%	0.4978%	0.5120%
DT	0.4941%	0.3028%	0.7041%	0.7614%	0.7025%	0.4637%	0.5714%
RFC	0.3415%	0.3352%	0.6889%	0.8445%	0.2731%	0.5704%	0.5089%
KNN	0.4227%	0.4020%	0.6799%	0.5495%	0.3958%	0.6818%	0.5220%
SVC-poly	0.3635%	0.3016%	0.7436%	0.5244%	0.4169%	0.6165%	0.4944%
LR	0.3837%	0.3443%	0.9345%	1.1220%	0.3039%	0.4987%	0.5979%
Average[15]	0.4091%	0.3218%	0.7318%	0.7818%	0.3988%	0.5541%	0.5329%[16]
benchmark	0.2646%	0.2058%	0.4488%	0.5259%	0.2496%	0.3945%	

Table 26 Daily profit efficiency with 2 labels / only activeness features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
GBM	0.4094%	0.3894%	0.2023%	0.3046%	0.3683%	0.4416%	0.3526%
SVC	0.2330%	0.3614%	0.3989%	0.4079%	0.5767%	0.3988%	0.3961%
DT	0.3217%	0.3424%	0.4549%	0.3765%	0.2334%	0.4415%	0.3617%
RFC	0.3815%	0.4345%	0.4530%	0.7004%	0.3327%	0.5100%	0.4687%
KNN	0.3110%	0.3183%	0.3514%	0.3623%	0.2963%	0.5591%	0.3664%
SVC-poly	0.2374%	0.3599%	0.2569%	0.3826%	0.2817%	0.3895%	0.3180%
LR	0.2882%	0.3030%	0.4000%	0.3024%	0.1511%	0.6092%	0.3423%
Average	0.3117%	0.3584%	0.3596%	0.4052%	0.3200%	0.4785%	0.3723%
benchmark	0.2071%	0.2332%	0.2888%	0.3501%	0.1629%	0.3936%	

Table 27 Daily profit efficiency with 3 labels / all features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
GBM	0.3400%	0.4901%	0.8058%	0.5839%	0.4531%	0.4772%	0.5250%
SVC	0.4398%	0.6163%	1.1913%	1.0507%	0.5244%	0.8820%	0.7841%
DT	0.5361%	0.5043%	1.1174%	1.1221%	0.3209%	0.8156%	0.7361%
RFC	0.3166%	0.3861%	0.7216%	0.7527%	0.4336%	0.5277%	0.5231%

[14] The Average column at the right of the table are the average ratio with respect to different stocks with same model.

[15] The Average lane at the bottom of the table are the average ratio with respect to same stock with different models.

[16] This cell means the average of all 42 tests with 6 stocks and 7 models. It reflects the quality of this feature set and label set.

KNN	0.5372%	0.8428%	1.2682%	0.6029%	0.5955%	0.9771%	0.8040%
SVC-poly	0.5067%	0.5746%	1.4916%	1.0327%	0.5512%	0.7739%	0.8218%
LR	0.5542%	0.5199%	0.9779%	0.9213%	0.5193%	0.6911%	0.6973%
Average	0.4615%	0.5620%	1.0820%	0.8666%	0.4854%	0.7349%	0.6987%
benchmark	0.2646%	0.2058%	0.4488%	0.5259%	0.2496%	0.3945%	

Table 28 Daily profit efficiency with 3 labels / only activeness features

	AAPL	GOOG	MSFT	NFLX	TSLA	AMZN	Average
GBM	0.4742%	0.4128%	0.4015%	0.5722%	0.2948%	0.4615%	0.4362%
SVC	0.4488%	0.4650%	0.4150%	0.9141%	0.3070%	0.6039%	0.5256%
DT	0.4439%	0.2699%	0.4127%	0.7059%	0.3442%	0.3446%	0.4202%
RFC	0.3221%	0.4767%	0.4212%	0.4835%	0.3665%	0.4945%	0.4274%
KNN	0.4381%	0.6473%	0.4769%	0.4353%	0.4666%	0.8088%	0.5455%
SVC-poly	0.3881%	0.3131%	0.4285%	0.6482%	0.6332%	0.7904%	0.5336%
LR	0.3374%	0.4416%	0.5646%	0.3436%	0.2619%	0.5516%	0.4168%
Average	0.4075%	0.4323%	0.4458%	0.5861%	0.3820%	0.5793%	0.4722%
benchmark	0.2071%	0.2332%	0.2888%	0.3501%	0.1629%	0.3936%	

The good performance in “daily profit efficiency” indicates that the predictions based on sentiment analysis can detect the growing day with a large increase and avoid dropping days with a large decrease.

6.3.2 Portfolio simulation

The “Portfolio” in the finance field corresponds to a combination of financial assets (stocks, foreigner exchange, bonds or even cash). When discussing stock investment strategy, the Portfolio means the stock assets arrangement that maximizes the expected profit while reducing the risk. Comparing with the trading strategy for single stock, the portfolio trading strategy increase the assets utilization rate because all investment is earning a profit (in expectation) while cash do not.

Comparing with the single stock trading strategy, the Portfolio strategy allows the investors to put their money into other investment opportunities to get expected profit when the algorithms choose not to own the stock for the next day.

The Portfolio trading strategy is illustrated as below.

- The trade decisions are made at the market close time of each day.
- In every trading day, all the money is used to purchase the stocks. The assets are arranged as follows:
- If none of the 6 stocks are predicted to grow in next day, the money will be used to averagely buy the same values (in the aspect of money instead of shares) of 6 stocks.
- If *X* stocks are predicted to grow in next day, then the money is arranged until each stock have $\frac{1}{X}$ of the total money. For instance, if stock AAPL, MSFT, and TSLA are predicted to grow for the next days, $\frac{1}{3}$ of the money will be used to buy AALP stock, $\frac{1}{3}$ will be used to buy MSFT and the same values of stocks for TSLA.

The experiments based on Portfolio trading strategy are also implemented. The results are shown in Figure 23 and Figure 24. The predictions with the KNN model, 3-labels, “all features” and the prediction with the KNN model, 3-labels, “only activeness features” are selected as the basis of the trading strategy.

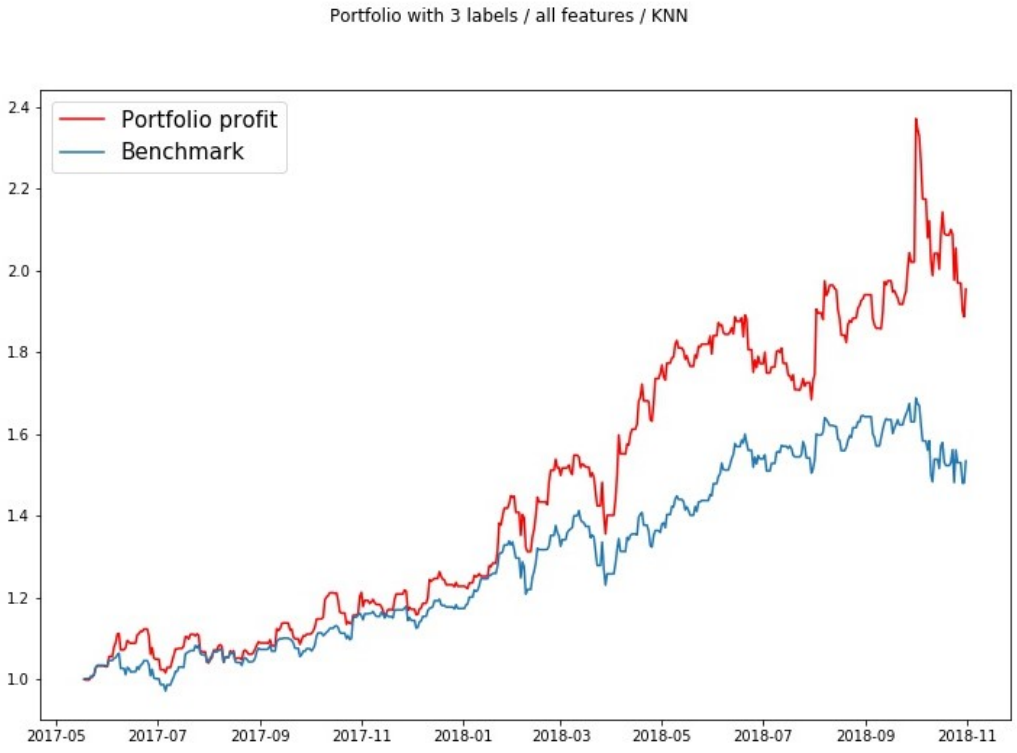


Figure 23 Portfolio profit with 3 labels / all features / KNN

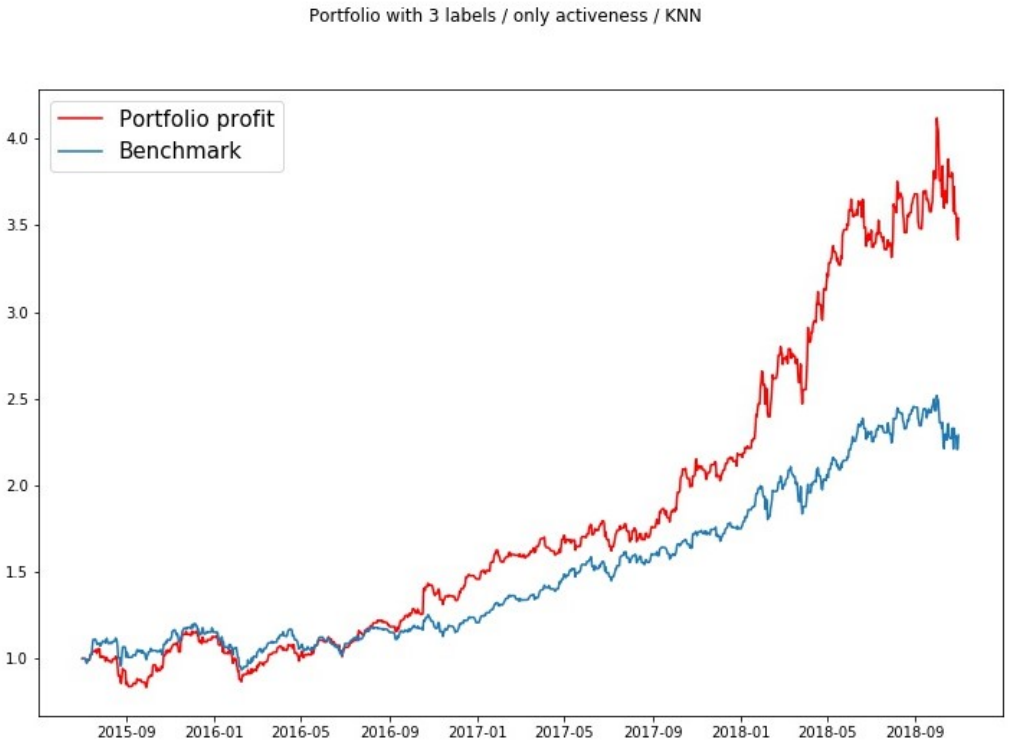


Figure 24 Portfolio profit with 3 labels / only activeness features / KNN

Form the figures, it is obvious that the profits based on the Portfolio trading strategy can profit more than benchmark strategy significantly. The improvement of algorithm

trading reveals the predictive power of Reddit text data. In addition, although the “only activeness” predictions have less “daily profit efficiency” than “all features”, but the performance in investment simulation is as good as the other one.

Comparing the results of two evaluation factors. Although the single stock trading strategy and portfolio trading strategy are based on the same predictions, they have totally different performances. The gap between them might come from two aspects. The first one is that prediction models can detect the day with large increasing or decreasing well. Catching the great growing days or avoiding collapsed days is the key to improving profit efficiency. The second reason might be the assets utilization rate. In the single stock trading strategy, the assets are in cash instead of being invested in other stocks when the algorithms decide not to buy the stock while the assets in B&H strategy and Portfolio strategies are making expected profits all the time. The high utilization rate carries a higher probability to get profit especially when the majority stocks are up trending (in the experiments above, 5 out of 6 stocks are up trending). It means, the bad performance in “total profit” might come from the assets utilization rate instead of prediction quality. Moreover, the stock TSLA, the only stock that can be considered as a “steady” trend, has the best prediction performance in “total profit” as well as the direction prediction accuracy.

In summary. The trading strategy based on predictions is profitable, and the features from sentiment analysis have a correlation with future price changes and are able to help predict the price movement.

7 CONCLUSION AND LIMITATION

7.1 Conclusions

In this thesis, sentiment analysis is applied to Reddit text data. Several processing algorithms and packages including the VADER sentiment package and SentiWordNet package are applied successfully. The sentiment analysis provides a set of sentiment features that explain the forum's emotions from many aspects. The descriptive statistics and predictive analysis have been applied to examine the relationship between Reddit forum text data and stock price changes. The descriptive statistics mainly consist of the Pearson correlation coefficient and correlation hypothesis tests. For descriptive statistics, 4 measurements are applied to evaluate the performance of predictions. The first factor is the prediction accuracy on price movement direction. The other three factors are based on the profits of algorithm trading simulations, respectively the “total profit with trading strategy for single stock”, “daily profit efficiency of trading strategy” and “total profit with Portfolio trading strategy”. Following the analysis mentioned above, the findings of this thesis can be concluded as below.

Firstly, the descriptive statistic results show the features that represent the activity degree of the Reddit forum have significant correlations with the future price changes in 1 to 4 days lags. Besides that, the sentiment features extracted via two sentiment packages do not show a consistent statistically significant linear correlation with future or historical price changes. However, the non-linear dependencies are exist based on the results of predictive analysis.

Secondly, the predictive analysis results have different performances under different measurements. With respect to the factors “prediction accuracy on price change direction” and “total profit with trading strategy for single stock”, almost all predictions based on different models and feature sets cannot outplay the B&H strategy. However, for the factors “daily profit efficiency of trading strategy” and “total profit with Portfolio

trading strategy", the predictions perform well that the vast majority of predictions with different models and feature sets are significantly better than B&H strategy.

Lastly, a Portfolio trading strategy that could profit more than B&H is designed. The trading strategy is fully based on the predictions with only the features extracted from the Reddit online forum. It proves that the prediction power of Reddit text data and provides the potential to make a profit in real stock markets.

The findings above show that the stock-related text in Reddit has a relationship with the stock price change. In addition, the investment simulations prove that the predictions in this thesis can help to make extra profit than the B&H strategy.

7.2 Limitation and future works

There are a couple of limitations in this study. Some of them lead to the potential of future research directions.

The first limitation is the data capacity. In this study, 6 stocks are used to validate the correlations. Although the time period is long, adding more stocks in different industries could make the experiment more reliable and persuasive. However, in Reddit, although many stocks are discussed and mentioned, the activity or data amounts of them are not enough for implementing any research on it. The selected 6 stocks in this thesis are the only stocks with enough text data. Fortunately, after 2017 or 2018, the activity and number of users of Reddit have increased very much. The higher data density and more active discussions make the text data in Reddit more feasible and suitable for further analysis. In addition, the data in other social media such as Twitter can also be used in stock prediction. Besides that, the diversity of stocks with different trends are needed. Based on the inference in section 6.3.2 and another work (Fu et al., 2013), the prediction methods have different performance when stocks have different trends (up, down, steady). Adding more stocks into validation might help to remove the bias and improving the prediction quality.

In addition, there are also limitations and shortcomings in research methods in this thesis. For the correlation hypothesis testing part, there is potential to apply some measurements that focus on multiple hypothesis situations like family-wise error rate or false discovery rate. Lastly, the investment simulations are originally designed to evaluate and compare the prediction performance of the models, so that the simulations are not fully based on real stock market rules. The stock trading processing, trading costs, and taxes are omitted. The un-preciseness makes that the trading strategies in this thesis may not be able to be applied or validated in real-world stock markets.

8 REFERENCES

Alambo, A., Gaur, M., Lokala, U., Kursuncu, U., Thirunarayan, K., Gyrard, A., . . . Pathak, J. (2019). Question answering for suicide risk assessment using reddit. Paper presented at the 468-473. doi:10.1109/ICOSC.2019.8665525

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge: MIT Press.

Ashley, C., & Tuten, T. (2015). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing, 32*(1), 15-27. doi:10.1002/mar.20761

Avdoulas, C., Bekiros, S., & Boubaker, S. (2018). Evolutionary-based return forecasting with nonlinear STAR models: Evidence from the eurozone peripheral stock markets. *Annals of Operations Research, 262*(2), 307-333. doi:10.1007/s10479-015-2078-z

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Paper presented at the *Lrec, , 10*(2010) 2200-2204.

Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems, 50*(4), 732-742. doi:10.1016/j.dss.2010.08.024

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit* " O'Reilly Media, Inc.".

Bollen, J., Mao, H., & Zeng, X. (2011). *Twitter mood predicts the stock market* doi://doi-org.libproxy.tuni.fi/10.1016/j.jocs.2010.12.007

Brett, E. I., Stevens, E. M., Wagener, T. L., Leavens, E. L. S., Morgan, T. L., Cotton, W. D., & Hébert, E. T. (2019). *A content analysis of JUUL discussions on social media: Using reddit to understand patterns and perceptions of JUUL use* doi://doi-org.libproxy.tuni.fi/10.1016/j.drugalcdep.2018.10.014

Brown, S. J., Goetzmann, W. N., & Kumar, A. (1998). The dow theory: William peter hamilton's track record reconsidered. *The Journal of Finance, 53*(4), 1311-1333.

Cergol, B., & Omladič, M. (2015). What can wikipedia and google tell us about stock prices under different market regimes? *Ars Mathematica Contemporanea, 9*(2), 301-320. doi:10.26493/1855-3974.561.37f

Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns: Method Matters—Evidence from the united states and italy. *Social Science Computer Review, 33*(1), 3-20. doi:10.1177/0894439314521983

Chan, E. (2013). *Algorithmic trading : Winning strategies and their rationale*. Somerset: John Wiley & Sons, Incorporated.

Chen, Y., Chen, Y., & Lu, C. L. (2017). Enhancement of stock market forecasting using an improved fundamental analysis-based approach. *Soft Computing, 21*(13), 3735-3757. doi:10.1007/s00500-016-2028-y

Contreras, I., Jiang, Y., Hidalgo, J. I., & Núñez-Letamendia, L. (2012a). Using a GPU-CPU architecture to speed up a GA-based real-time system for trading the stock market. *Soft Computing, 16*(2), 203-215. doi:10.1007/s00500-011-0714-3

Contreras, I., Jiang, Y., Hidalgo, J. I., & Núñez-Letamendia, L. (2012b). Using a GPU-CPU architecture to speed up a GA-based real-time system for trading the stock market. *Soft Computing, 16*(2), 203-215. doi:10.1007/s00500-011-0714-3

COOPER, M. J., GULEN, H., & OVTCHINNIKOV, A. V. (2010). Corporate political contributions and stock returns. *The Journal of Finance, 65*(2), 687-724. doi:10.1111/j.1540-6261.2009.01548.x

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. doi:10.1007/BF00994018

Cosma, S.Advanced data analysis from an elementary point of view. Retrieved from http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/

D. T. Tran, A. Iosifidis, J. Kanniainen, & M. Gabbouj. (2019). *Temporal attention-augmented bilinear network for financial time-series data analysis* doi:10.1109/TNNLS.2018.2869225

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy, 98*(4), 703-738.

Desmet, B., & Hoste, V. (2013). *Emotion detection in suicide notes* doi://doi-org.libproxy.tuni.fi/10.1016/j.eswa.2013.05.050

Dijck, J. v. (2013). *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199970773.001.0001

Evans, C., Pappas, K., & Xhafa, F. (2013). *Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation* doi:https://doi.org/10.1016/j.mcm.2013.02.002

Fama, E. F. (1965). Random walks in stock market prices. *Financial Analysts Journal, 21*(5), 55-59. doi:10.2469/faj.v21.n5.55

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383-417. doi:10.2307/2325486

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data* Cambridge University Press.

Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. Paper presented at the 1072-1081. doi:10.1109/HICSS.2016.137

Fu, T., Chung, F., & Chung, C. (2013). Adopting genetic algorithms for technical analysis and portfolio management. *Computers and Mathematics with Applications, 66*(10), 1743-1757. doi:10.1016/j.camwa.2013.08.012

FUNG, S. Y. K., SU, L. (., & ZHU, X. (. (2010). Price divergence from fundamental value and the value relevance of accounting information. *Contemporary Accounting Research, 27*(3), 829-854. doi:10.1111/j.1911-3846.2010.01028.x

Guo, X., Lai, T. L., Shek, H., & Wong, S. P. (2016). *Quantitative trading : Algorithms, analytics, data, models, optimization*. Boca Raton: CRC Press LLC.

H. Sul, A. R. Dennis, & L. I. Yuan. (2014). Trading on twitter: The financial information content of emotion in social media. Paper presented at the *2014 47th Hawaii International Conference on System Sciences,* 806-815. doi:10.1109/HICSS.2014.107

Hagenau, M., Liebmann, M., & Neumann, D. (2013). *Automated news reading: Stock price prediction based on financial news using context-capturing features* doi://doi-org.libproxy.tuni.fi/10.1016/j.dss.2013.02.006

Hootsuite Media, I. (2019). The global state of digital in 2019 report. Retrieved from https://hootsuite.com/pages/digital-in-2019

Hu, Y., Feng, B., Zhang, X., Ngai, E. W. T., & Liu, M. (2015). *Stock trading rule discovery with an evolutionary trend following model* doi://doi-org.libproxy.tuni.fi/10.1016/j.eswa.2014.07.059

Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., & Liu, M. (2015). *Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review* doi://doi-org.libproxy.tuni.fi/10.1016/j.asoc.2015.07.008

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Paper presented at the *Eighth International AAAI Conference on Weblogs and Social Media,*

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR), 47*(4), 1-38. doi:10.1145/2771588

J. Bollen, & H. Mao. (2011). Twitter mood as a stock market predictor. *Computer, 44*(10), 91-94. doi:10.1109/MC.2011.323

Jeong, C. Y., Lee, S. T., & Lim, J. (2019). Information security breaches and IT security investments: Impacts on competitors. *Information & Management, 56*(5), 681-695. doi:10.1016/j.im.2018.11.003

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access, 6*, 23253-23260. doi:10.1109/ACCESS.2017.2776930

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons, 53*(1), 68. doi:10.1016/j.bushor.2009.09.003

Kaucic, M. (2010). *Investment using evolutionary learning methods and technical rules* doi://doi-org.libproxy.tuni.fi/10.1016/j.ejor.2010.07.008

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2014). *Text mining for market prediction: A systematic review* doi://doi-org.libproxy.tuni.fi/10.1016/j.eswa.2014.06.009

Lee, C. J., & Andrade, E. B. (2015). Fear, excitement, and financial risk-taking. *Cognition and Emotion, 29*(1), 178-187. doi:10.1080/02699931.2014.898611

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. Paper presented at the *Fourth International AAAI Conference on Weblogs and Social Media,*

Li, Q., Li, P., Wang, T., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences, 278*, 826-840. doi:10.1016/j.ins.2014.03.096

Li, Q., Wang, T., Gong, Q., Chen, Y., Lin, Z., & Song, S. (2014). *Media-aware quantitative trading based on public web information*

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14-23. doi:10.1016/j.knosys.2014.04.022

Lin, C., Weng, R. C., & Keerthi, S. S. (2007). Trust region newton methods for large-scale logistic regression. Paper presented at the 561–568. doi:10.1145/1273496.1273567

Lin, P., & Chen, J. (2007). *FuzzyTree crossover for multi-valued stock valuation* doi://doi-org.libproxy.tuni.fi/10.1016/j.ins.2006.08.017

Lo, A. W. (2005). Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. *Journal of Investment Consulting, 7*(2), 21-44.

Lv, D., Huang, Z., Li, M., & Xiang, Y. (2019). Selection of the optimal trading model for stock investment in different industries. *PloS One, 14*(2), e0212137. doi:10.1371/journal.pone.0212137

Mankiw, N. G. (1998). *Principles of economics*. Fort Worth (Tex.): Dryden Press.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM, 38*(11), 39-41.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). *Sentiment analysis on social media for stock movement prediction* doi://doi-org.libproxy.tuni.fi/10.1016/j.eswa.2015.07.052

Nofsinger, J. R. (2005). Social mood and financial economics. *Journal of Behavioral Finance, 6*(3), 144-160. doi:10.1207/s15427579jpfm0603_4

Nyce, C., & Cpcu, A. (2007). Predictive analytics white paper. *American Institute for CPCU.Insurance Institute of America,* , 9-10.

Obar, J. A., & Wildman, S. (2015). Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy, 39*(9), 745-750. doi:10.1016/j.telpol.2015.07.014

Oberlechner, T., & Hocking, S. (2004). *Information sources, news, and rumors in financial markets: Insights into the foreign exchange market* doi://doi-org.libproxy.tuni.fi/10.1016/S0167-4870(02)00189-7

Olson, R. S., & Neal, Z. P. (2015). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science, 1*, e4. doi:10.7717/peerj-cs.4

P. Nousi, A. Tsantekidis, N. Passalis, A. Ntakaris, J. Kanniainen, A. Tefas, . . . A. Iosifidis. (2019). *Machine learning for forecasting mid-price movements using limit order book data* doi:10.1109/ACCESS.2019.2916793

Pandrekar, S., Chen, X., Gopalkrishna, G., Srivastava, A., Saltz, M., Saltz, J., & Wang, F. (2018). Social media based analysis of opioid epidemic using reddit. *AMIA ...Annual Symposium Proceedings.AMIA Symposium, 2018*, 867-876.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Reddit. (2017). Press - reddit. Retrieved from https://www.redditinc.com/press

Robertson, C., Geva, S., & Wolff, R. (2006). What types of events provide the strongest evidence that the stock market is affected by company specific news?

Rosati, P., Cummins, M., Deeney, P., Gogolin, F., van der Werff, L., & Lynn, T. (2017). The effect of data breach announcements beyond the stock price: Empirical evidence on market activity. *International Review of Financial Analysis, 49*, 146-154. doi:10.1016/j.irfa.2017.01.001

scikit-learn.Scikit-learn images. Retrieved from https://scikit-learn.org/stable/_images/sphx_glr_plot_svm_margin_001.png

Sezer, O. B., Ozbayoglu, M., & Dogdu, E. (2017). A deep neural-network based stock trading system based on evolutionary optimized technical analysis parameters. *Procedia Computer Science, 114*, 473-480. doi:10.1016/j.procs.2017.09.031

Shen, K., & Tzeng, G. (2015). *Combined soft computing model for value stock selection based on fundamental analysis* doi://doi-org.libproxy.tuni.fi/10.1016/j.asoc.2015.07.030

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems, 85*, 74-83. doi:10.1016/j.dss.2016.03.001

Siering, M., Deokar, A. V., & Janze, C. (2018). *Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews* doi://doi-org.libproxy.tuni.fi/10.1016/j.dss.2018.01.002

Siikanen, M., Baltakys, K., Kanniainen, J., Vatrapu, R., Mukkamala, R., & Hussain, A. (2018). *Facebook drives behavior of passive households in stock markets* doi:https://doi-org.libproxy.tuni.fi/10.1016/j.frl.2018.03.020

Strauß, N., Vliegenthart, R., & Verhoeven, P. (2016a). *Lagging behind? emotions in newspaper articles and stock market prices in the netherlands* doi://doi-org.libproxy.tuni.fi/10.1016/j.pubrev.2016.03.010

Strauß, N., Vliegenthart, R., & Verhoeven, P. (2016b). Lagging behind? emotions in newspaper articles and stock market prices in the netherlands. *Public Relations Review, 42*(4), 548-555. doi:10.1016/j.pubrev.2016.03.010

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2014). Trading on twitter: The financial information content of emotion in social media. Paper presented at the 806-815. doi:10.1109/HICSS.2014.107

Telang, R., & Wattal, S. (2007). An empirical analysis of the impact of software vulnerability announcements on firm stock price. *IEEE Transactions on Software Engineering, 33*(8), 544-557. doi:10.1109/TSE.2007.70712

Urquhart, A., & Hudson, R. (2013). *Efficient or adaptive markets? evidence from major stock markets using very long run historic data* doi://doi-org.libproxy.tuni.fi/10.1016/j.irfa.2013.03.005

Wang, J., & Leu, J. (1996). Stock market trend prediction using ARIMA-based neural networks. Paper presented at the *, 4* 216-2165 vol.4. doi:10.1109/ICNN.1996.549236

Wei, P., & Wang, N. (2016). Wikipedia and stock return: Wikipedia usage pattern helps to predict the individual stock movement. Paper presented at the 591–594. doi:10.1145/2872518.2890089

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* doi://doi.org/10.1016/j.iref.2012.07.016

Yu, L., Wu, J., Chang, P., & Chu, H. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems, 41*, 89-97. doi:10.1016/j.knosys.2013.01.001

Yu, Y., Duan, W., & Cao, Q. (2013a). *The impact of social and conventional media on firm equity value: A sentiment analysis approach* doi://doi-org.libproxy.tuni.fi/10.1016/j.dss.2012.12.028

Yu, Y., Duan, W., & Cao, Q. (2013b). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919-926. doi:10.1016/j.dss.2012.12.028

Yudong, Z., & Lenan, W. (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. *Expert Systems with Applications, 36*(5), 8849-8854. doi:10.1016/j.eswa.2008.11.028

Zayats, V., & Ostendorf, M. (2017). Conversation modeling on reddit using a graph-structured LSTM.

9 APPENDIX

Table 29 Abbreviations in Literature review tables

ARMA	AutoRegression Moving Average
ATR	Average True Range
B&H	Buy and Hold Strategy
BMA	Bayesian Model Averaging
BoF	Bag of Feature
BOOST	Boosting method
BPNN	Back Propagation Neural Network
CART	Classification And Regression Tree
CCI	Commodity Channel Index
CCR	Charnes, Cooper and Rhodes
CMA	Colume Moving Average
D/BE	Debt over Book Equity
DBN	Deep Belief Network
DEMATEL	Decision-making trial and evaluation laboratory
DRSA	Dominance-based rough set approach
DT	Decision Tree
EMV	Ease of Movement Value
FCA	Formal concept analysis
GA	Genetic Algorithm
GP	Genetic Programming
GRU	Gated Recurrent Unit
HPR	holding-period-return
IBCO	Improved Bacterial Chemotaxis Optimization
K&D	stochastic (K&D) indicator
LOB	Limit Order Book
LR	Logistic Regression
LSTAR	Logistic STAR
LSTM	Long Short Term Memory
MA	Moving Average
MACD	Moving Average Convergence Divergence
MCSDA	Multilinear Class-Specific Discriminant Analysis
MDA	Multilinear Discriminant Analysis
MFI	Money Flow Index
MLP	Multilayer Perceptron
MTM	Momentum
NB	Naïve Bayes Model
NBoF	Neural Bag of Feature
NN	Neural Network

OBV	On Balance Volume
OGrossProfit	Operational gross profit/total revenue
OProfit	Operational profit/total revenue
P/B	Price-to-Book
PCF	Price-to-Cash flow
PVC	Plurality Voting Committee
RF (RFC)	Random Forest (Classifier)
RNN	Recurrent Neural Network
ROA	Return on asset
ROC	Rate of Change
RSI	Relative Strength Index
RSV	Raw Stochastic Value
SAE	Stacked Auto-Encoders
SETAR	Self-Exciting TAR
SG	Growth in sales
SLFN	Single Hidden Layer Feedforward Neural Networks
STAR	Smooth Transition AR
STC	Stochastic Oscillator
SVM	Support Vector Machine
TAR	Threshold Autoregressive
TOG	turnover growth
TRIX	Triple Exponentially Smoothed Moving Average
W%R	William % R
XCS	eXtended Classifier Systems
XGB	eXtreme Gradient Boosting algorithm