UNIVERSITY OF CALIFORNIA
SANTA CRUZ

STOCK PREDICTION BY ANALYZING FINANCIAL NEWS SENTIMENT AND INVESTOR MOOD OF SOCIAL MEDIA

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Jia Deng

December 2020

The Thesis of Jia Deng
is approved:

_____
Professor Yi Zhang, Chair

_____
Professor Marilyn Walker

_____
Professor Yang Liu

_____
Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

STOCK PREDICTION BY ANALYZING FINANCIAL NEWS SENTIMENT AND INVESTOR MOOD OF SOCIAL MEDIA

Jamie Deng

Abstract

Stock prediction is a difficult task. Recently many studies focus on Natural Language Processing methods to draw information from texts and perform forecasting. This paper purpose to build a stock prediction system which utilize textual data from multiple resources such as financial news and social media feeds. The system applies different NLP methods like VADER analyzer and Word2vec representation, to extract sentiment scores and investor mood from texts. Then it combines the data with historical prices and volumes and uses Random Forest and SVM algorithms to train models, which make predictions on short-term stock price movements. Experiment results show good accuracy and F1 scores for some of the stocks and for the combined model of all stocks. The highest accuracy is for Apple shares, 75.68%. I also simulate two trading strategies based on the news sentiment and investor mood indicators respectively. They both outperform simple buy-and-hold strategy.

1 Introduction

Financial markets, especially the stock markets, play an very important role in world economy. Predicting stock price movements is a very difficult task, since there are lots of factors affecting the market. Researchers and industry practitioners have been studying the predictability of stock markets for decades [5]. Classic economic theory like the efficient market hypothesis (EMH) [16] assumes that share prices already reflect all available information, thus it's impossible for a strategy to consistently outperform the market. However, behavior finance [6] suggests some degree of predictability. Investors' mood and overconfidence could cause prices to deviate systematically from their fundamental values [19]. Many empirical studies also show that stock prices are to some extent predictable [5].

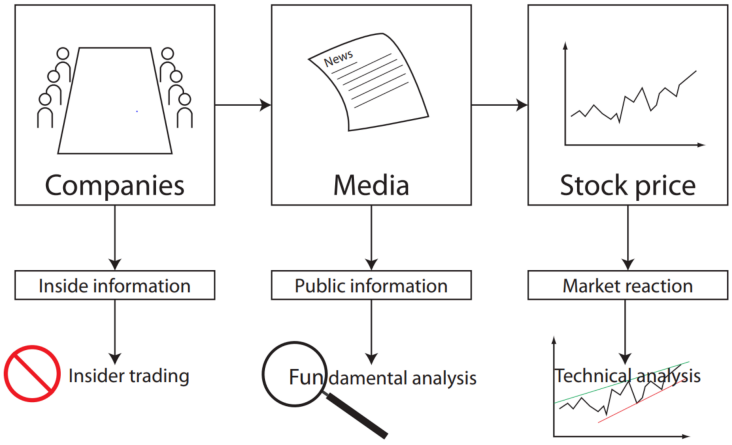

Figure 1: Stock analysis approaches [27]

There are two common approaches (Figure 1) to analyze and predict financial market [4]: *Fundamental analysis* focus on economic factors and financial information that influence asset price movement. *Technical analysis* models historical price as time series. Traditionally statistical models are used to model time series. They assume linear process, but financial time series are complex and nonlinear. In recent years, machine learning (ML) and its application on financial market forecasting has gain much attentions [13]. They are without statistical assumptions,

nonlinear and relatively successful [4]. In addition to research on financial time series, recent studies also incorporate mining info from text format such as financial news, reports, and social media to improve the accuracy of ML predictions.

This project will focus on the fundamental approach, which based on textual data of sentiments of financial news and investor mood of social media to predict future price movements. I propose a stock prediction system, which use Natural Language Processing (NLP), such as VADER [10] sentiment analyzer and Word2vec representation, to extract information from textual data. Then it employs Random Forest and SVM algorithms to make predictions on future price movements. The experiment results show some very good accuracy and F1 scores. I also simulate two trading strategies based on those sentiment and mood indicators. They both make very good profits compared to simple buy-and-hold strategy.

2 Related Work

In practice, traders have been using textual data to improve financial market modeling for a long time [28]. They traditionally rely on different kinds of texts such as financial reports, press releases, equity and market research report, and news articles to make better informed investment decisions. In recent year, social media platforms like Twitter and StockTwits have produced an large amount of user contents. Financial news and social media texts are readily accessible and frequently updated, which make them potentially suitable for automatic processing and analyzing. I will focus on these two types of textual data.

[22] survey the methods of mining financial news, and claim that news are one of the most effective sources that affect the market. [20] propose a systematic approach to a predict directional movements of a currency pair by mining news headlines. Their system is an online method, which updates the model with new data instances therefore is robust to concept drift. [15] use attention based recurrent neural

networks to extract rich semantic features from news headlines. They claim that using headlines produce better results than using the articles. The authors employ bi-directional long short term memory networks (LSTM) to encode news text and capture context information. They also apply self attention mechanism to distribute attention on most relative words, news and days. [7] utilize event representation learning on news corpus with pre-trained event embedding. Their models are enhanced with additional common sense knowledge of intent and sentiment of the event participants. Those research show very competitive results with high prediction accuracies.

There are also many studies focus on social media contents for stock prediction. [3] investigate and find that public mood data collected from Twitter feeds do have correlation with DJIA. They claim that prediction accuracy can be significantly improved by adding certain public mood indicator such as "clam". Their study draw much attention of other researchers. Sentiment analysis of Twitter data is applied for predicating stock market movements [24]. The authors use two different textual representations: Word2vec, N-gram, to perform sentiment analysis on tweets. Then they apply machine learning method to analyze the correlation between stock movements and sentiments in tweets. The studies show very high accuracies.

Some research focus on StockTwits, which is a social media platform dedicated to investing community. Sentiment lexicon and sentiment-oriented word vector are learned form StockTwits [14]. The authors claim that investor sentiment indicators can predict stock market, and show that the domain-specific lexicon and sentiment word embedding outperform general methods. Other research involving StockTwits also produce very promising results. Sentiment scores are calculated by analysis of feeds through SVM [2]. Feature selection method is applied to identify the relevant terms in twits and decision tree is built to determine the trading decisions based on terms. Then a trading strategy is constructed to evaluate the profitability of the term trading [1].

3 Methods

The aim of this project is to build a stock prediction system (Figure 2) which predicts short term stock price movements based on multiple sources of data. The price can either go up or down. Therefore it's a binary classification problem. The system use NLP methods to extract sentiment information from news headlines, and investor mood from social media platform. Then it combines these two types of textual data with prices and volume time series to train machine learning models, that will make predictions of future price trends based on current available information.

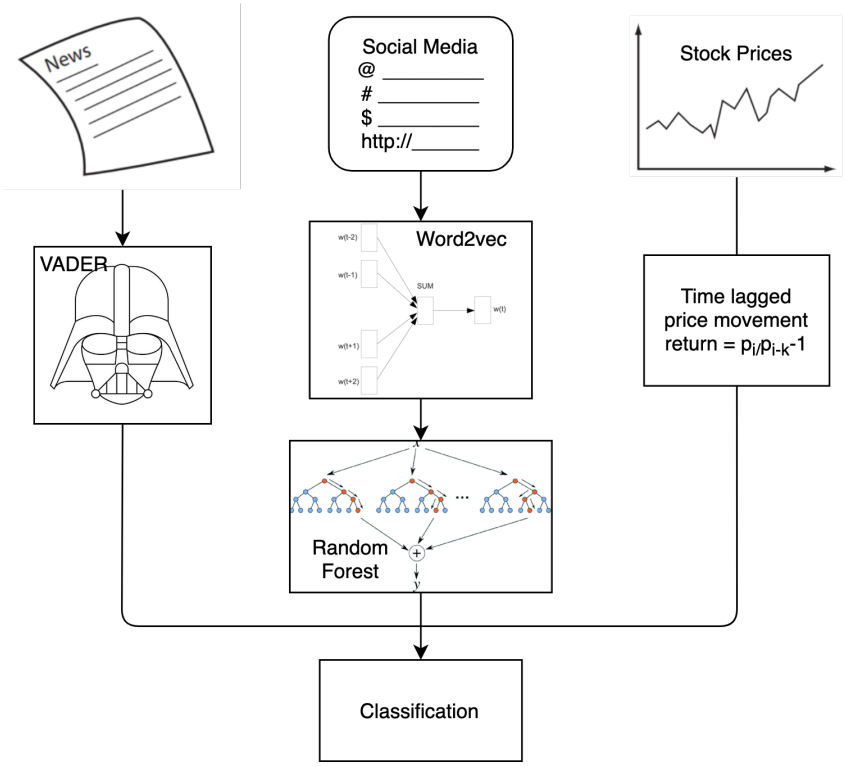

Figure 2: Flow chart of the proposed system

3.1 Datasets

For experiment purpose, I select the five largest companies (Figure 3) from S&P 500 index of the US stock market. They happen to be the so-called big techs: Apple, Microsoft, Amazon, Google (owned by Alphabet), and Facebook. Tesla will join

the index in Dec 2020. These companies are all in the technology sector and heavily traded. Fluctuations in their prices largely affect the market index, because they are so big. Apple has reached 2 trillion dollars market value. I have three types of data: historical prices and volume data, financial news headlines, and social media twits. News headlines are selected instead of full content of the articles, since headlines produce better results [15].

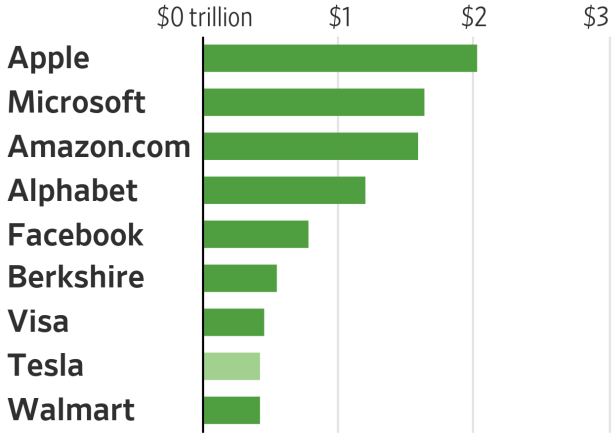

Figure 3: Market values of the largest S&P 500 companies [26]

3.1.1 Data Collection

The prices and volume data are retrieved from Yahoo Finance. For the textual data, financial news headlines are fetched from NASDAQ website, and social media data are obtained from StockTwits, which is a social media platform designed for sharing ideas between investors, traders, and researchers.

For each stock, I collect the data specifically dedicated to it. Each stock will have two to three years of data, depending on the availability of financial news in the website's archive. All data instances are dated. Each day involves one instance of price and volume, but incorporates multiple news headlines, even thousands or tens of thousands social media twits. For example, Apple is very popular among traders, there are 2 years of price data, but 9900 news headlines, and 660,000 twits.

3.1.2 Data Processing

Although financial news data are unlabeled, social media twits are partially user annotated. When user post a twit, they can choose to provide a label: bullish or bearish, which represents their opinion about whether the stock is going up or down. For example, 46% twits of Apple are user annotated, in which 69% are bullish. Therefore I can utilize a supervised machine learning algorithm to train a model on these labeled data, and use the model to make predictions of investors' mood on the unlabeled part of the dataset.

In order to analyze trends of the stocks, I need to process the data both before and after applying NLP methods. News headlines usually comply with standard format, but I still need to do some simple pre-processing. For example, stopword and special character removal are applied to remove words and symbols that do not express any emotions. StockTwits feeds are tokenized, during which punctuations, emoticons, symbols, and urls are removed. After sentiment and mood information are extracted, they need to be aggregated based on dates. There are also dates discrepancies between textual and prices data, since stocks are not traded during weekends and holidays. Reverting sentiment and mood data back to trade days solves this issue. Calculating daily returns (Equation 1) from the prices and converting positive and negative returns into Up and Down labels produce the target for our classification task. In this equation, k is the time lag.

$$\text{daily return} = \frac{P_i}{P_{i-k}} - 1 \tag{1}$$

3.2 NLP for textual data

For news headlines and social media feeds, I apply different NLP methods to extract information from the textual format. Since news headlines data are unlabeled, unsupervised learning method is required. However, the StockTwits feeds are

partially annotated. Users have the option to choose the tag of bullish or bearish to express their anticipation of the future stock prices. This provides the opportunity for application of machine learning method to learn on labelled data and classify the unlabeled ones.

3.2.1 News Sentiment Analysis

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a crowd sourced lexicon and rule-based tool for sentiment analysis. The authors of this method claim that [10] VADER outperforms human readers and generalize better than benchmarks like SentiWordNet, General Inquirer, and machine learning techniques based on Naive Bayes and Support Vector Machine. VADER analyzer is very useful for short text. It produces both polarity (positive/negative) and intensity (strength) of emotion. The model relies on a dictionary that maps lexical features to emotion intensities, i.e. sentiment scores. Summing up the intensity of each word in the text generates the compound score of a text.

VADER relies on a dictionary which maps words and expressions to valence scores. It's tuned to sentiment expression in microblogs. Creating a thorough dictionary is labour-intensive and time-consuming. Therefore the authors build on widely used dictionaries to categorize words and lexical features. They expand the capacity of the features using the Wisdom of the Crowd approach, which relies on the collective opinion of a group of individuals rather than that of a single expert. Human raters use heuristics to analyze the sentiment. It helps the model to make valid estimate of the sentiment score of the texts. By summing up the valence scores of each word, the analyzer computes the compound score, then adjust and normalized the score to be between $(-1, 1)$. I use the compound scores for stock prediction. Each day's sentiment score is aggregated by combining the extreme values of the day's compound scores.

3.2.2 StockTwits Mood Learning

Since StockTwits feeds are partially labelled, it's possible to train a machine learning model on the labelled data instances and classify the investor mood of the rest unlabelled ones. Such machine learning algorithm requires features extraction of the twits. Word2vec [17] is a technique to construct word embedding from textual datasets. It can be learned by using two methods, Skip Gram and Common Bag of Words. The Word2vec algorithm uses a Neural Network to learn word associations from the corpus of text. It takes the corpus as input and generate a high dimensional vector space. Each word in the corpus can be represented as a vector. Relationship between the words is retained in their vector format. Each twit's vector representation is a weighted sum of every word it contains. The weights are computed by using TFIDF (term frequency-inverse document frequency) of the corpus.

After converting twits to numerical features, the labelled data is divided into training set and test set. A Random Forest model is trained on the training set. The model performs very well in classifying the test dataset. For the twits about Apple, it produces 73.19% accuracy and 83.21% F1 score (Table 1). According to [24], the rate of agreement among humans readers on the sentiment of a text, is between 70% and 79%. They consider results above 70% are accurate in most cases. Then this model is ready to predict on the unlabelled twits. The stock prediction system combines the predicted labels with user annotated ones for model training.

	Accuracy	Precision	Recall	F1
AAPL	0.7319	0.7355	0.958	0.8321
AMZN	0.74	0.7633	0.9411	0.8429
FB	0.7253	0.7436	0.9327	0.8257
GOOG	0.7677	0.781	0.9712	0.8657
MSFT	0.8695	0.8791	0.986	0.9295

Table 1: Results of investor mood prediction

3.3 Model Training

Once I have all the data and resolve the date differences, it's split into 85% training set and the rest is test set. The stock prediction system trains Random Forest models on the training set. SVM algorithm is also used for comparison purpose. According to [8], Random Forest is the one of the best performing machine learning algorithms. [4] suggest that Artificial Neural Networks are very popular and effective in forecasting financial market. But they also present some research which show that RF outperform Neural Networks. [18] find that RF outperform other methods (not including NN). Other studies [5, 9] show that Deep Learning especially LSTM outperforms RF with a small margin. RF is chosen for the system because of its efficiency and good performance.

The system computes the daily returns from the prices time series, then relabels any positive/negative returns as Up/Down. [3] and [24] both suggest time lag $k = 3$ days between data instances and prediction. The sentiment scores are aggregated by combining the extreme compound scores of each day. For investor mood scores, the Bullish/Bearish labels are converted into $1/-1$ and added for each day. The system use the changes of sentiment and investor mood scores from the previous day as feature inputs. Since the datasets are unbalanced with more Ups than Downs, the algorithm adjusts with balanced sampling. Number of trees in RF is set to be 128 considering the trade-off between accuracy and computational costs, since no significant gain in accuracy of RF when the number of trees goes beyond 128 [23]. The system firstly trains on data for each stock to make prediction on each one. Then it also aggregates the data of all stocks to train a combined model.

4 Result and Discussion

Experiment results show that for some stock, e.g. Apple, the performance of stock prediction system is better than that of the others. And with the combined model,

prediction accuracies for all stocks improve. I also formulate two trading strategies based on news sentiment and investor mood indicators respectively. They both perform very well.

4.1 Performance

Experiment results of the stock prediction system by using RF are shown in Table 2 and Table 3. The single models are trained on each stock's own data. Combined model is trained on aggregated data of all stocks. Both approaches utilize Random Forest algorithm. It's clear that the combined model outperforms single models, especially for the stock of Apple, Amazon, and Google. Their performance improve a lot. The prediction on Apple price movements has the best accuracy of 75.68% and the best F1 score. However for FB and GOOG, the improvements are not significant. It seems like the system produces better results with more data to train. Since those stocks are in the same sector and correlated, the aggregation makes sense. But more historical data may not enhance the results, as market conditions and investor perceptions change over time and concept drift sets in.

Stock	Accuracy	Precision	Recall	F1
AAPL	0.5714	0.6452	0.6667	0.6557
AMZN	0.5106	0.5484	0.6538	0.5965
FB	0.5309	0.5814	0.5556	0.5682
GOOG	0.5241	0.6322	0.5978	0.6145
MSFT	0.5814	0.6923	0.6429	0.6667

Table 2: Prediction results of single models, RF

Stock	Accuracy	Precision	Recall	F1
AAPL	0.7568	0.7463	0.9804	0.8475
AMZN	0.6197	0.6197	1.0	0.7652
FB	0.5702	0.6437	0.7273	0.6829
GOOG	0.6147	0.6129	1.0	0.76
MSFT	0.6047	0.7432	0.6322	0.6832

Table 3: Prediction results of combined model, RF

For comparison purpose, I train the models by using SVM algorithm and produce similar results (Figure 4). The combined model significantly outperforms single models, especially for AAPL, AMZN, and FB. Except for GOOG, the accuracy is worse. For MSFT, the accuracy and F1 scores slightly increase. The SVM algorithm produce better scores for AMZN, FB, and MSFT. But RF performs better on AAPL and GOOG. For both algorithms, the highest accuracy and F1 scores belong to AAPL stock.

SVM	Single Models			
	Accuracy	Precision	Recall	F1
AAPL	0.6122	0.6122	1.0	0.7595
AMZN	0.5532	0.5532	1.0	0.7123
FB	0.5679	0.5641	0.9778	0.7154
GOOG	0.6345	0.6345	1.0	0.7764
MSFT	0.6512	0.6512	1.0	0.7887

	Combined Model			
	Accuracy	Precision	Recall	F1
AAPL	0.7297	0.7313	0.9608	0.8305
AMZN	0.6338	0.6324	0.9773	0.7679
FB	0.6364	0.6364	1.0	0.7778
GOOG	0.6101	0.6101	1.0	0.7578
MSFT	0.6822	0.6917	0.954	0.8019

Table 4: Results of SVM algorithm

4.2 Trading Strategies

In order to test the effectiveness of sentiment and mood indicators, I construct two simple trading strategies that utilize those two indicators respectively. They perform very well compared to a simple buy-and-hold strategy. I simulate these strategies on Apple stock during one year period from 2019-10-16 to 2020-10-16. In all trading simulations, commission is set to be zero because major brokers in the US no longer charge commissions any more. The first strategy is based on news sentiment indicator. It buys 100 shares of Apple when the sentiment score

increases by 0.2 points from the previous day. And it sells 100 shares when the sentiment score decreases by 0.2 points. The strategy only allows a buy order if previous position has been settled.

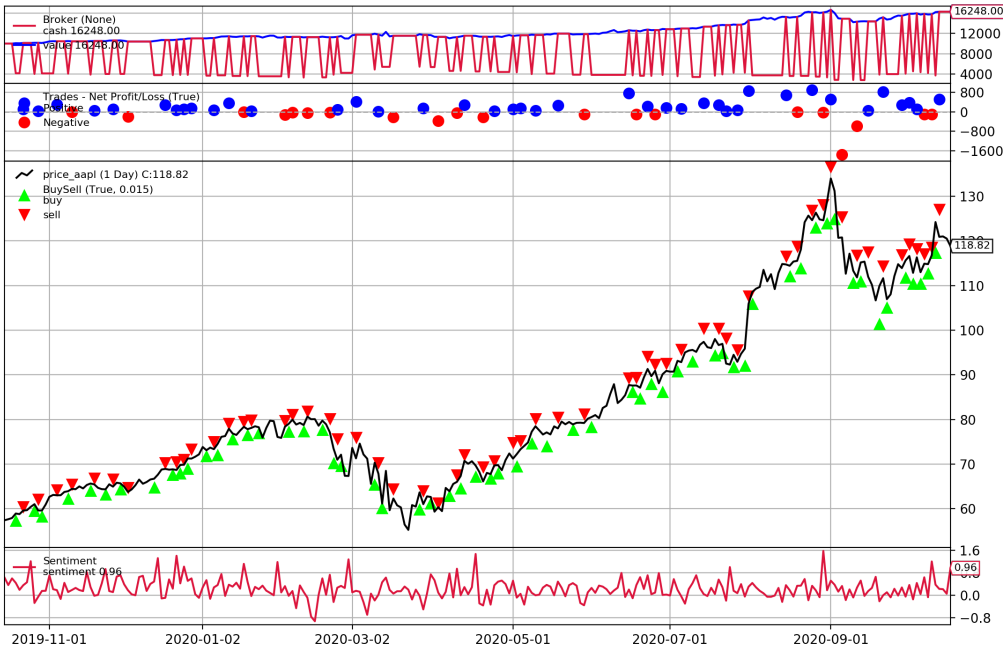

Figure 4: Performance of strategy based on sentiment indicator

Figure 4 shows the results of the sentiment based strategy. Assume the starting position is $10,000 cash, after one year the profit ratio is 62.48%. The second strategy is based on investor mood indicator. It buys 100 shares when aggregate mood number increases by 20%, and sells when the number decreases by 20%. Results (Figure 5) show that it produces 83.8% profit ratio. A simple buy-and-hold strategy is to buy 100 shares at the beginning of this time period, then hold the shares till the end. This would generate 61.32% profit. The caveat is that stock prices of Apple almost doubled during this time period. That's the reason why these profit margins are so high. If the overall trend of share price goes down, these simple strategies may not work. Then more sophisticated methods are required.

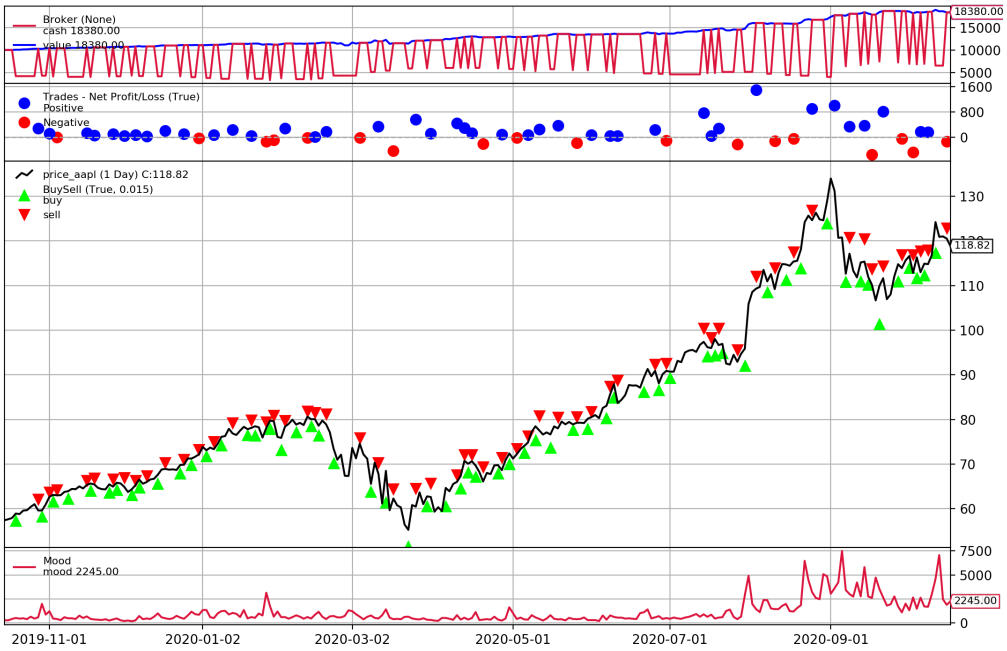

Figure 5: Performance of strategy based on mood indicator

Figure 4 demonstrates that the sentiment indicator successfully identifies the peak price which is around 133, and the strategy sells at that point. On the other hand, Figure 5 shows that the mood indicator picks up the lowest price point at 56, then 100 shares are bought. It seems like the sentiment and mood indicators do have some predictive power over the price movements of Apple.

4.3 Discussion

In this project, I draw data from multiple sources including historical prices and volumes, and textual data such as financial news, and StockTwits feeds. The stock prediction system employs different NLP methods and extracts features from the news headlines and social media feeds. It also applies different aggregation methods for news and social media feature values. By combining those features with historical time series, the system produce good results compared to the baseline models (Table 5). In addition, I construct two trading strategies based on news sentiment and investor mood indicators respectively. The related works usually

draw data from single textual source. They usually select sentiment scores as feature and use simple average to aggregate the values. Few of them build strategy to act on signals extracted from texts.

Baselines	Data	Method	Performance	Trading strategy
Hu et al. 2018 [11]	Chinese stocks, News	Hybrid Attention Networks, self paced learning	Accuracy 48%	buy top 10 stocks with highest up prob, best annual return 1.5
Nguyen et al. 2015 [21]	18 stocks, message board texts	Sentiment analysis	Acc 54.41%	No
Pagolu et al. 2016 [24]	Microsoft stock, Tweets, partially manual labeled	supervised sentiment analysis, Word2vec	Acc 69% Logistic regr, 71.8% SVM, good correlation	No
Prosky 2017 [25]	Tech stocks, News	mood analysis, event extraction	Acc average 50% highest 0.722	No
Xu & Cohen 2018 [29]	88 stocks, Twitter feeds	latent variables, neural variational inference	Acc 58.23% highest	No
Liu 2018 [15]	SP 500, News headlines, pre-trained word embedding	Attention based LSTM	Acc 63%, highest 72%	No
Ding et al. 2019 [7]	NYT corpus, pre-trained event embedding	event repr learning with intent, emotion embedding	Acc 68.5% highest	No

Table 5: Methods and results of baseline models

The proposed stock prediction system produces average accuracy of 63.32% and highest accuracy score of 75.68% by using Random Forest. It also produces average accuracy 65.85% and highest accuracy 72.97% by using SVM algorithm. These scores are competitive compared to the baseline models. The results show that financial news and social media texts do have predictive power over short-term share price movements, as suggested by some of the related studies. The simulated trading strategies also perform very well with high annual returns. Furthermore, the system produces much higher recall scores than accuracy scores. It means that the system is better at predicting the Up direction, even after it's being adjusted with balanced sampling.

5 Conclusion and Future Work

I build a stock prediction system which utilize textual data from multiple resources such as financial news and social media twits. The system employs different Natural Language Processing methods like VADER sentiment analyzer and Word2vec representation, to extract sentiment scores and investor mood information from the texts. Then it combines the data with historical prices and volumes to train machine learning models, which make predictions on future stock price movements. The experiment results show very good accuracy and F1 scores for some of the stocks and for the combined model. The highest accuracy is 75.68% for Apple stock, which is competitive compared to state-of-the-art methods [24, 15, 7, 11, 12, 25, 29]. I also simulate two trading strategies based on those sentiment and mood indicators. They both make very good profits compared to simple buy-and-hold strategy.

There could be a lot of possible future work on the topic of stock prediction. For example, We could draw data from more diversified sources, apply different NLP and aggregation methods on different types of texts, expand the experiment to more stocks from other sectors, and develop more advanced trading strategies.

References

[1] Al Nasseri, A., Tucker, A., and de Cesare, S. Quantifying stocktwits semantic terms trading behavior in financial markets: An effective application of decision tree algorithms. *Expert systems with applications 42*, 23 (2015), 9192–9210.

[2] Batra, R., and Daudpota, S. M. Integrating stocktwits with sentiment analysis for better prediction of stock price movement. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (2018), IEEE, pp. 1–5.

[3] Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of computational science 2*, 1 (2011), 1–8.

[4] Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., and Oliveira, A. L. Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications 55* (2016), 194–211.

[5] Chong, E., Han, C., and Park, F. C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications 83* (2017), 187–205.

[6] DellaVigna, S. Psychology and economics: Evidence from the field. *Journal of Economic literature 47*, 2 (2009), 315–72.

[7] Ding, X., Liao, K., Liu, T., Li, Z., and Duan, J. Event representation learning enhanced with external commonsense knowledge. *arXiv preprint arXiv:1909.05190* (2019).

[8] Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research 15*, 1 (2014), 3133–3181.

[9] Fischer, T., and Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research 270*, 2 (2018), 654–669.

[10] Gilbert, E., and Hutto, C. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf* (2014), vol. 81, p. 82.

[11] Hu, Z., Liu, W., Bian, J., Liu, X., and Liu, T.-Y. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining* (2018), pp. 261–269.

[12] Kim, R., So, C. H., Jeong, M., Lee, S., Kim, J., and Kang, J. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999* (2019).

[13] Krauss, C., Do, X. A., and Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research 259*, 2 (2017), 689–702.

[14] Li, Q., and Shah, S. Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (2017), pp. 301–310.

[15] Liu, H. Leveraging financial news for stock trend prediction with attention-based recurrent neural network. *arXiv preprint arXiv:1811.06173* (2018).

[16] Malkiel, B. G. The efficient market hypothesis and its critics. *Journal of economic perspectives 17*, 1 (2003), 59–82.

[17] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] Milosevic, N. Equity forecast: Predicting long term stock price movement using machine learning. *arXiv preprint arXiv:1603.00751* (2016).

[19] Nardo, M., Petracco-Giudici, M., and Naltsidis, M. Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys 30*, 2 (2016), 356–369.

[20] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications 42*, 1 (2015), 306–324.

[21] Nguyen, T. H., Shirai, K., and Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications 42*, 24 (2015), 9603–9611.

[22] Nikfarjam, A., Emadzadeh, E., and Muthaiyah, S. Text mining approaches for stock market prediction. In *2010 The 2nd international conference on computer and automation engineering (ICCAE)* (2010), vol. 4, IEEE, pp. 256–260.

[23] Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (2012), Springer, pp. 154–168.

[24] Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (2016), IEEE, pp. 1345–1350.

[25] Prosky, J., Song, X., Tan, A., and Zhao, M. Sentiment predictability for stocks. *arXiv preprint arXiv:1712.05785* (2017).

[26] Ramkumar, A. What tesla's addition to the s&p 500 means for investors. https://howtoassetmanage.com/what-teslas-addition-to-the-sp-500-means-for-investors/, Nov. 2020.

[27] Singh, S. Machine learning for stock market investment - an introduction. https://ieee.nitk.ac.in/blog/machine-learning-for-stock-market-investment/, Nov. 2017.

[28] Xing, F. Z., Cambria, E., and Welsch, R. E. Natural language based financial forecasting: a survey. *Artificial Intelligence Review 50*, 1 (2018), 49–73.

[29] Xu, Y., and Cohen, S. B. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 1970–1979.