ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

Prediction of DJIA based on Machine Learning and Natural Language Processing

Shuxiao Zhou[1], Huiyi Yuan[2], Yuxuan Zhao[3], Chengxiang Wen[4]

[1]Department of Economics and Management, Wuhan University, 430072, Wuhan, China;

[2]School of Business, University of Connecticut, 06269, Storrs, United States;

[3]Department of Mathematical Science, Nankai University, 300071, Tianjin, China;

[4]Suizhou No.2 high school, 441300, Suizhou, China.

Abstract

This work harnessed news headlines in a machine leanring and natural language processing framework to predict the DJIA up and down moves. Several finding will be showed as follow 1) News headlines contribute to the accuracy of predictions. 2) The Twitter sentiment does not drive the accuracy of stock market predictions. 3) There are differences in the results of the stock market forecast using different models.

Keywords

Machine Learning; Natural Language Processing; SVM.

1. Introduction

The stock market influences the social life in many aspects and has become a very popular research field in academia. The related research of the stock price forecast has attracted extensive attention. Early stock market forecasts were based on a well-known hypothesis, the efficient market hypothesis [1]. The efficient market hypothesis holds that stock prices are a function of information and rational expectations. The current share price gives an almost immediate picture of the company's prospects [2]. This means that all public information about a company, including its historical price, is absorbed into its current share price. Therefore, changes in stock prices are closely related to the release of new information. Malkiel [2], in his influential work A Random Walk Down Wall Street, argued that stock price history could not accurately predict stock prices. For this reason, Malkiel argues that stock prices are best described by a statistical process called a "random walk," in which daily deviations from the center are random and unpredictable. While the efficient market hypothesis is popular among finance scholars, its critics point out that actual market experience differs from the unpredictability it implies. Warren Buffett, one of the most famous and successful investors, refuted the efficient market hypothesis in a 1984 speech at Columbia University. Gradually, people began to shift their attention away from the efficient market hypothesis and look for other more accurate and efficient methods of forecasting stock prices.

One attempt is the addition of financial news in a quantitative trading system. The prediction of stock markets leveraging the news or personal insights requires extracting valuable characteristics from texts using text mining, classifying positive and negative influence on stock price and simulating stock price returns [3]. A research was conducted to analyze and extract such information, and derive numerical indicators from financial text [4]. However, the news, in some cases, presents both positive and negative aspects of the stock markets in somewhat a neutral tone, which makes it difficult to figure out the truth behind such news [5]. Some text mining systems have been proposed in order to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

overcome such limitations, where unstructured big data are gathered, parsed, tagged, analyzed and converted to opinions suitable for making stock market predictions [3].

There have been many research studies aimed at identifying that relationship or predicting stock market movements using news analysis. Various types of news extracted on the web [6], such as emails and blogs, have been found to closely predict stock market behavior. Sehgal and Song [5] gathered, analyzed and extracted individual investors' opinions disclosed on the web, analyzed their sentiments, calculated their author's reliability, and predicted the stock values of three companies via machine learning. Mittermayer and Knolmayer [7] leveraged the stock prediction system News CASTS, analyzed media news on specific companies, and experimented with a comparison between news and stock price flows. Lavrenko et al. [8] described a unique system for predicting trends in stock prices based on the content of stories that precede the trends, which can identify the news stories that are highly indicative of future trends.

Another attempt is the sentiment analysis of social networks. In recent years, social media has been penetrating into every aspect of people's daily life. At the same time, with the continuous development of big data technology and artificial intelligence, researchers have made it possible to track and predict users' behaviors on social media platforms. Many researchers began to use group emotional data on stock financial markets and construct the corresponding forecasting model in stock [9]. Relevant psychological research points out that human emotions also play an important role in decision-making. Behavioral finance provides further evidence that financial decisions are largely driven by emotions. Therefore, this paper infers that public sentiment has a certain driving effect on the stock market. [10,11]. Twitter is a large and popular social networking service that allows millions of users to share in real-time about events worthy of wide attention and to express public opinion. Therefore, a large number of useful public sentiment texts were collected on Twitter and conduct in-depth research on the trend of public sentiment on stock prices and how to improve accuracy [12].

Support vector machine (SVM) is a two-class classification model whose basic model is defined as the linear classifier with the largest spacing on the feature space. The learning strategy is to maximize the spacing, which ultimately translates into the solution of a convex quadratic planning problem [13]. So far, support vector machines have been very widely used and have solved many practical problems. As the theoretical knowledge of the support vector machine has improved, it has become more stable. Based on the features of textual data learning analysis, fully automated learning can be ideally achieved by improving support vector machines without the demand for manual parameter tuning [14]. In fact, in the realm of text analysis, the performance of many learning algorithms degrades as the vector of words transformed by document parameters increases. However, support vector machines are immune to the high dimensionality of the text space [14], which results in higher accuracy [15]. Enable clustering analysis of news and other texts by implementing clustering algorithms by SVM analysis of similarity features [16]. Schumaker and Chen [17] used various textual based representations in a machine learning framework to predict asset price directional moves.

This research uses historical news headlines from Reddit WorldNews Channel and Dow Jones Industrial Average (DJIA) to integrate features from behavioral finance and text mining, seeking to empirically determine whether valuable characteristics extracted from text can be used to predict future directional moves of the DJIA. This work applies natural language processing methods to news headlines, aggregates the mood states of tweets and uses four traditional machine learning model to predict directional moves of the DJIA.

2. Data

In this research, there are two datasets were used for the stock prediction model which are from an online data community Kaggle.com. The first one is the News and DJIA dataset, and the second one is the Twitter sentiment dataset.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

2.1 News and DJIA

There are two feeds of data provided in this dataset: News data and Stock data [18]. For the news data, some historical news headlines were selected from Reddit WorldNews Channel. They are ranked by Reddit users' votes, and only the top 25 headlines are considered for a single date. The timeline is from 2008-06-08 to 2016-07-01. There are two columns provided for this part. The first column is the "date", and the second column is the "news headlines". All news are ranked from top to bottom based on popularity (news that are most discussed). For each day in the date range, the top 25 ranked news headlines are provided.

The other dimension of the Kaggle dataset is stock price data. The DJIA was applied and the timeline is between 2008-08-08 to 2016-07-01. The stock price information was downloaded directly from Yahoo Finance. Table 1 below shows the summary statistics for the stock data.

Table 1. Summary statistics for stock price data

	Mean	Std. Deviation	Quantiles				
			Min	25%	50%	75%	Max
Open	13.5k	3.14k	6.55k	10.9k	13k	16.5k	18.3k
High	13.5k	3.14k	6.71k	11k	13.1k	16.6k	18.4k
Low	13.4k	3.15k	6.47k	10.8k	13k	16.4k	18.3k
Close	13.5k	3.14k	6.55k	10.9k	13k	16.5k	18.3k
Volume	163m	93.9m	8.41m	100.0m	135m	193m	675m
Adj Close	13.5k	3.14k	6.55k	10.9k	13k	16.5k	18.3k

Note: This table provides the statistics data (Mean, Std Deviation, Min, and Quantiles) for different status of the stock market.

The data was splited up into train data and test data. For the test one, the date ranged from 2008-08-08 to 2014-12-31 and for testing is from 2015-01-02 to 2016-07-01, which results in the typical 80% and 20% ratio.

2.2 Twitter sentiment dataset

This dataset contains the data provided by Lachanski and Pav (2017) to reproduce their paper: *"Twitter Mood Predicts the Stock Market"*. Since they are unable to share the raw tweets, this dataset contains summary statistics for the Twitter data and the DJIA (2018). The data contains Twitter sentiment data aligned with returns of the DJIA and GSPC (S&P 500 index). There are 11 fields in this dataset (Table 2). The timeline is between 2007-07-19 to 2008-12-31.

3. Methodology

The NLP approach based traditional machine learning focuses on pre-processing and feature extraction of text. Then the processed text is vectorized and finally the training dataset is modeled by a common machine learning algorithm. In traditional machine learning algorithms, the quality of feature extraction of the text has a significant impact on the accuracy of text classification.

In this paper, the data before Dec 1st, 2018 is e the training set and the remaining is the Test Set, which follows roughly a 80%/20% split. There are two labels for stock movement: "1" when DJIA Adj Close value rose; "0" when DJIA Adj Close value decreased, as shown in Figure 1. For both Training Set and Test Set, the news headlines ranging from "Top 1" to "Top 25" are required to be joined together for further processing. Term frequency–inverse document frequency (TF-IDF) is applied as a feature extraction algorithm, which identifies words in a collection of documents that are useful for determining the topic of each document. A word has a high TF-IDF score in a document if it appears in relatively few documents, but appears in this one, and when it appears in a document it

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

tends to appear many times. News data is first put into training models to determine how the prediction matches the real stock movement. Then, both news and twitter sentiment is leveraged to replicate the same steps. In the news data, only the effect of the particular mood, "calm", is considered, since Bollen, Mao and Zeng (BMZ) [9] find that only "calm" has reasonable predictive power for changes in the DJIA.

Table 2. Fields Explanation

Colums'name	Explanation
date	The date of the observed Twitter sentiment data. The data a
tone	The tone sentiment series.
calm	The calm sentiment series.
tone_Z10	The normalized tone variable, using k=10. This is essentially a centered Z-score of tone over 21
tone_Z1	The normalized tone variable, using k=1. This is essentially a centered Z-score of tone over 3 (!) observations.
DJI	The closing value of the DJI index for the given day. Will take value of NA when the market is closed. Early closes are treated as normal market days. The DJI data are source from Yahoo finance via the quantmod package.
DJI*volume*k	The volume of the DJI index for the given day, in thousands. This may be useful for data QA, for example. The volume is NA when the market is closed.
DJI*forward*return	This is a one market period relative 'return' of the DJI index. A value of 0.01, for example, corresponds to a 1% increase in the DJI index. This is a forward return, meaning it is the return from the close of the given day to the close of the next market day, and could be approximately captured by a long holder in the index. (Were that possible; an index is not an ETF.) The return is NA when the market is closed. Note that there is an overlap in information between the Twitter sentiment series and the forward return: the sentiment data is only observable just before midnight EST, while one would have to invest just before 4PM EST to capture the forward return.
GSPC	The closing value of the GSPC index for the given day. Will take value of NA when the market is closed. Early closes are treated as normal market days. The GSPC data are source from Yahoo finance via the quantmod package. While the original paper does not make claims regarding Twitter mood and the GSPC series, it represents broad market returns and makes a suitable target for a putative forecast of the market by sentiment.
GSPC*volume*k	The volume of the GSPC index for the given day, in thousands.
GSPC*forward*return	The one market period relative 'return' of the GSPC index. This is a forward_return, meaning it is the return from the close of the given day to the close of the next market day. The return is NA when the market is closed.

Note: This table explains what each column in the data set means

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

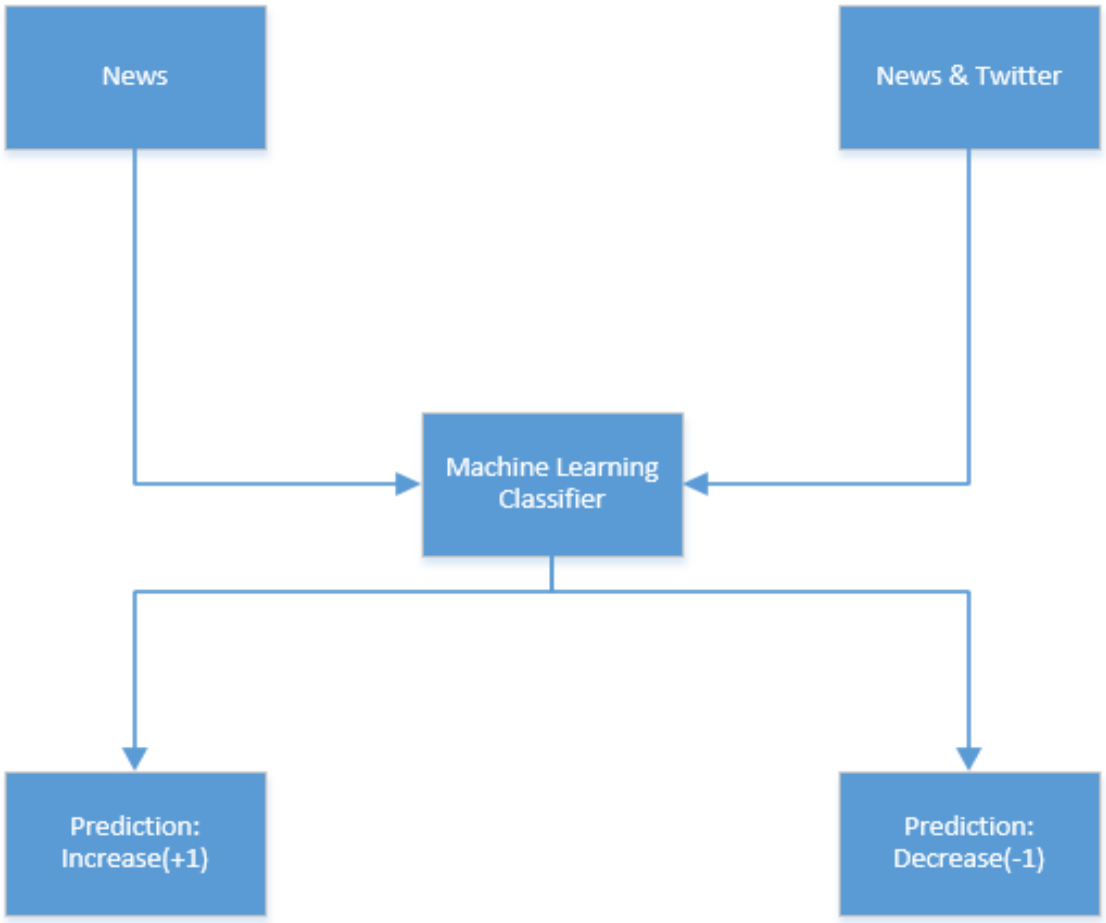

Figure 1. Workflow of our models

Four well-known traditional machine learning classifiers are adopted to train and test a model in order to predict the stock movement. The machine learning classifiers are listed below

(1) **Support Vector Machine**: a method of classification and regression which constructs a hyperplane or set of hyperplanes in a high-dimensional or infinite-dimensional space [19].

(2) **Logistic Regression**: a statistical model uses a logistic function to model a binary dependent variable.

(3) **Naïve Bayes**: Bayesian parameter estimation based on some prior distribution [20].

(4) **Random Forest**: an ensemble method that constructs a multitude of decision tree at training time [21].

Finally, the models are evaluated by accuracy, precision, recall and F-score.

4. Result

After training, four models built by four algorithms which are Logic Regression, Naïve Bayes, Random Forest and Support Vector Machine are set up. Adjustments have been made to the vectorization of news headlines of each model and after testing them separately with testing dataset and Compared with the actual value, they all got solid results after optimization of this work.

This work maily used Tf-idf method to do the feature extraction, and Logic Regression got about 0.47 F-score, while Naive Bayes got about 0.78 F-score. Random Forest got different F-score every time the code was run, and mostly higher than the F-score of Support Vector Machine, which is around 0.53. As can be seen, Naïve Bayes got the best performance when considering F-score. The table 3 below will show the precision.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

Table 3. Precision

	Logic Regression	Naïve Bayes	Random Forest	SVM
Accuracy	0.59	0.77	0.68	0.68
Precision	0.67	0.75	0.83	1
Recall	0.36	0.82	0.45	0.36
F-score	0.47	0.78	0.59	0.53

5. Conclusion

Predicting stock prices, which are affected by many factors, is a challenging task. This paper proposes a new method that applies the natural language processing method and the establishment of four models to news headlines and integrates Twitter sentiment to predict the stock price trend. The contributions of this study can be summarized as follows. First, a general stock price forecasting framework is implemented. Second, while considering the news factor in previous studies, this study proposes a method to use Twitter sentiment to assist and improve the accuracy of stock market prediction. Finally, the dataset of Twitter was combined with the news. Then the work used four models to conduct correlation calculation and test model performances against each other respectively. To test whether Twitter's sentiment helped improve the accuracy of stock predictions, this work removed the Twitter data and tested it again using the headline data. By experimenting with news headlines and twitter's sentiment, the following conclusions were drawn:

(1) News headlines do contribute to the accuracy of forecasts. At the stock, industry, and index levels, models with news headline data sets showed relative accuracy in both validation and independent test data sets.

(2) The Twitter sentiment didn't drive or improve the accuracy of stock market predictions. Simply focusing on the positive and negative dimensions does not lead to useful predictions. Models using emotional polarity performed poorly on all tests.

(3) There are slight differences in the results of the stock market forecast using four different models.

There are still some limitations to this paper and there are some solutions to improve them in the future. The extraction of news events through news headlines did not verify the difference between news headlines and news articles on the degree of impact on stock prices. In addition, the Twitter sentiment data collected is within one year and the data set does not have completeness and representativeness. For future work, expanding the dataset will be considered. As time goes on and more data becomes available, it is believed that the results of the model will improve. Furthermore, some significant financial events are not taken into account, which requires further improvement and expansion of event classification in the future. The follow-up work will further optimize the algorithm to improve the stock market prediction by joining the neural learning network and focus on data mining similarity calculation, correlation analysis, clustering techniques, and natural language processing of text summarization, text generation, and other in-depth analysis.

This research paper applies data mining, machine learning, and natural language processing techniques to the stock market analysis. After the improvement of the models and of the stock price prediction accuracy, it is hoped that the research results of this article can be widely used in relevant financial practical fields.

References

[1] Fama, E. F.. (1970). Efficient capital markets: a review of theory and empirical work. The Journal of Finance, 25, 383-417.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0053

[2] Malkiel, B. G.. (1990). A random walk down wall street: including a life-cycle guide to personal investing. Ww Norton & Company, 40(17), 1566.

[3] Kim, Y., Jeong, S. R., & Ghani, I.. (2014). Text Opinion Mining to Analyze News for Stock Market Prediction.

[4] Gidofalvi, G., & G Gidófalvi. (2001). Using news articles to predict stock price movements. department of computer science & engineering.

[5] Sehgal, V., & Song, C. (2007). SOPS: Stock Prediction Using Web Sentiment. Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), 21–26.

[6] Das, S. R. , & Chen, M. Y. . (2008). Yahoo! for amazon: sentiment extraction from small talk on the web. Operations Research, 48(6), 601-602.

[7] Mittermayer, M. A., & Knolmayer, G. F.. (2006). NewsCATS: A News Categorization and Trading System. IEEE International Conference on Data Mining.

[8] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J.. (2000). Language Models for Financial News Recommendation. ACM, 389-396.

[9] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

[10] Nofsinger, & John, R.. (2005). Social mood and financial economics. Journal of Behavioral Finance, 6(3), 144-160.

[11] Stanton, Steven, J., Reeck, Crystal, Huettel, & Scott, A., et al. (2014). Effects of induced moods on economic choices. Judgment & Decision Making.

[12] Houlihan, P., & Creamer, G. G.. (2019). Leveraging social media to predict continuation and reversal in asset prices. Computational Economics(3).

[13] Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567.

[14] Joachims, T.. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. Proc. Conference on Machine Learning. Springer, Berlin, Heidelberg.

[15] Cooley, R. (1999). Classification of News Stories Using Support Vector Machines. in IJCAI'99 Workshop on Text Mining.

[16] Finley, T., & Joachims, T. (2005). Supervised clustering with support vector machines. Proceedings of the 22nd International Conference on Machine Learning - ICML '05, 217–224.

[17] Schumaker, R. P., & Chen, H.. (2009). A quantitative stock prediction system based on financial news. Information Processing & Management, 45(5), 571-583.

[18] Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [Date You Retrieved This Data] from https://www.kaggle.com/aaron7sun/stocknews.

[19] Cortes, C., & Vapnik, V.. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[20] Russell, S. J., & Norvig, P. N. (2009). Artificial Intelligence: A Modern Approach. Prentice Hall. applied mechanics & materials.

[21] Ho, T. K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282 vol.1.