Analysing effect of news polarity on stock market prediction: a machine learning approach

Conference Paper · November 2021

DOI: 10.1109/IKT54664.2021.9685403

4 authors, including:

Golshid Ranjbaran
Islamic Azad University Tehran Science and Research Branch
9 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Mohammad-Shahram Moin
Iran Telecommunication Research Center
82 PUBLICATIONS **809** CITATIONS
SEE PROFILE

Abbas Koochari
Islamic Azad University Tehran Science and Research Branch
5 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Analyzing effect of news polarity on stock market prediction: a machine learning approach

1st Golshid Ranjbaran
Faculty of Electrical and Computer Science
Islamic Azad University Science and Research Branch
Tehran, Iran
golshid.ranjbaran@srbiau.ac.ir

2nd Mohammad-Shahram Moin
ICT Research Institute
Tehran,Iran
moin@itrc.ac.ir

3rd Sasan H Alizadeh
IT Research Faculty
ICT Research Institute
Tehran, Iran
s.alizadeh@itrc.ac.ir

4th Abbas Koochari
Faculty of Electrical and Computer Science
Islamic Azad University Science and Research Branch
Tehran, Iran
koochari@srbiau.ac.ir

***Abstract*— In finance, the stock market and its trends are volatile in nature. In the stock market, which is dynamic, complex, nonlinear and non-parametric, accurate forecasting is crucial for trading strategy. This need attracted researchers to detect fluctuations and to predict the next move. It is assumed that news articles affect the stock market. In this work, non-measurable data like financial news headlines has been transferred into the measurable data. We investigated the relationship between news and their impact on stock prices. To show this relationship, we applied the sentiment analysis data and the price difference between the day before the news was published and the day of the news to the classic machine learning models such as SVR, BayesianRidge, LASSO, Decision tree and Random forest. The observations showed that SVM performs well in all tests. The prediction error in this model is 0.28, which is much less than that of the random news tagging. Also based on our tests, using a computer for tagging is as good as manual tagging.**

Keywords—News, stock price prediction, sentiment analysis, machine learning

I. INTRODUCTION

In finance, the stock market and its trends are volatile in nature. This attracts researchers to detect fluctuations for predicting the next move. Investors and market analysts study market behavior and plan their buying or selling strategies accordingly. Because the stock market produces a lot of data every day, it is very difficult to take into account all the current and past information to predict the future trend of stocks. There are basically two ways to predict market trends. One is technical analysis and the other is fundamental analysis. Technical analysis considers the price and volume of past trades to predict future trends. On the other hand, a fundamental analysis of a business involves the analysis of its financial data to gain insight into the performance of the company. The effectiveness of technical and fundamental analysis is disputed by the efficient market hypothesis, which states that stock market prices are fundamentally unpredictable.

Fundamental analysis is the science of evaluating economic, financial, and other variables about an asset that can determine its current and probable future value. In other words, in fundamental analysis, the analyst examines and analyzes the conditions of a project or company from different aspects and follows its important news and events. Analyzing all aspects and factors affecting a project or company in order to identify its intrinsic value and potential is known as the definition of fundamental analysis.

It is not easy to examine all the factors influencing a company's stock, especially for small investors who are looking for short-term investment. These people deicide to buy or sell stocks based on the feeling that the news evokes in them. For example, if the news of the acceptance of Bitcoin by Amazon is published. Given that Amazon is a large company that millions of people buy or sell from this company every day, the person assumes this news positive and will buy the share of Amazon, and thus a positive news can lead to growth the price [1].

Stock market forecasting has long been an active area of research. The efficient market hypothesis (EMH) states that stock market prices are heavily influenced by new information and follow a random pattern. This hypothesis is known as behavioral economics and states that public mood and market performance are interrelated. The idea is that when people are happy and optimistic, there is the potential for increased investment, which in turn improves stock market performance.

News agencies as one of the sources that people and investors visit every day and this visit, unlike in the past, which required buying a newspaper or visiting a website, is easily possible by just subscribing to a social network because all the news agencies have their own page on social networks and reflect the news of the day in it. Therefore, the news published in these news agencies easily affects the public opinion and the opinions of investors. Examining this news and the feeling it

978-1-6654-0849-3/21/$31.00 ©2021 IEEE

evokes in the reader can be considered as a very important factor in predicting the stock price [2].

II. RELATED WORKS

Predicting stock price trends is an attractive area of research because more accurate forecasts are directly related to higher stock returns. Thus, in recent years, considerable efforts have been made to develop models that can predict the future trend of a particular stock or general market. Most existing techniques use technical analysis. Some researchers have shown that there is a strong relationship between news about a company and its stock price. The following is a discussion of previous research on sentiment analysis of news and stock prices.

Lee uses the Daily News Sentiment Index (DNSI) and Google Trends data in coronavirus searches to examine the initial impact of COVID-19 sentiment on the US stock market. The goal is to examine the relationship between the COVID-19 sentiment and the 11 selected US stock market indices over a period of time. Any public positive or negative feelings about the stock market crisis can have an impact on investors' decisions in the stock markets. The results show the distinct effects of COVID-19 sensation in different industries and separate them into different correlation groups[3].

The goal is to create an effective model for predicting stock market trends with small error and improve forecasting accuracy. This model is based on the analysis of sentiments and stock market prices and is designed using two methods K-NN and simple Bayesian algorithm. Khedr separates the model into two stages, the first is to determine the positive or negative polarity of the news using a simple Bayesian algorithm, the second is to output the first as input along with past share prices to predict future stock trends using The K-NN algorithm combines[4].

Kaliani et al. Project use data such as financial news articles about a company and predict the future trend of its stock by classifying news sentiment, assuming that news articles affect the stock market. This is an attempt to examine the relationship between news and stock trends. To this end, they used a dictionary-based approach. Dictionaries of positive and negative words are created using words that carry specific public and financial sentiment. Based on these data, they implemented classification models. The results show that Random Forest (RF) and Support Vector Machine (SVM) work well in all experiments[5].

Research by Ding et al. Suggests that news can affect stock market behavior, and that yesterday's news can affect daily stock price changes. They tried to extract a procedure for extracting information from news headlines using a process called open information extraction. The result shows that their method works well. They proved with experiments that news headlines should be sufficient for textual features and thus improve the share price forecast[6].

III. RESEARCH METHODOLOGGY

A. Dataset

The data set used in this article includes 20015 news items that have been collected from 240 news websites, including the New York Times, Washington Post, etc., which are available to the public through link[1]. This dataset also includes visual data, but we only used the textual data. It is a combination of all the news that is happening around the world, and in this article we only needed news that was somehow related to the stock under review, so by applying the filter only the news that in the headline or in the body of the news contained a keyword of the stock, was selected. For example, to select news related to Apple company keywords such as 'apple', 'iPhone', 'aapl' are applied to the main data and the related news was extracted.

B. Preprosessing

The mentioned data only includes the news that has been collected from different news agencies and the stock price was received separately from yahoofinence.com on the day of publishing the news. Based on the price of the day before (before the news was published) and the day the news was published, the price difference created by the news was calculated.

Textual data is unstructured data. Therefore, we cannot provide raw data as input to the regression model. In this stage of preprocessing, first all numbers, symbols, abbreviations, ineffective words, etc. were removed, and then TextBlob was used as a tool to extract the sentiment of each news. This tool gives each news a number between 0 and 1.

C. Research methodology

In this research, we have tried to answer the following questions:

Does the news act credibly and affect the stock market properly?

Can more accurate predictions be made with positive news or negative news?

How much better is predictive accuracy in manual tagging than computer tagging ?

Figure 1 shows a model for using the news and sentiments we have created to predict stock price changes Then the news related to the stock and the changes that this news has led to, are extracted., The amount of positive and negative sentiments of the published news has also been estimated These sentiments have been used to predict stock price changes . In this way, the regression model first learns and predicts price changes using positive sentiments, and then we repeated the process with negative sentiments.

As most research shows, SVM, Random Forest and Naive Bayes work well in regression models [7,8]. We separated 80% of the data for the training phase and 20% for the test phase. The simulation is implemented in the Python programming environment using the scikit-learn library. Therefore, we have examined all three algorithms along with several other

[1] https://drive.google.com/open?id=0B3e3qZpPtccsMFo5bk9Ib3VCc2c

algorithms to predict the price change that has occurred. The outputs are given in the continuation of the article.

In the data surveyed, people manually tagged the news positively and negatively and we called it 'manual tagging'. We compared these tags with sentiment analysis tags using the TextBlob library and called it 'computer tagging', because manual tagging is costly, and much more news can be tagged if computers can be used for tagging. This tagging can also be done online and quickly. Therefore, it is possible to act very quickly to predict changes.

We used positive and negative sentiments separately as well as in combination to predict price changes that occurred.

Also to answer the question of whether the news really affects the stock market or not, we compared the accuracy of the predict with random data tagging and news sentiment analysis data so that we can answer the question clearly.

For the record, random data tagging is data that is generated randomly and sentiment analysis data are output of the news from the TextBlob library.

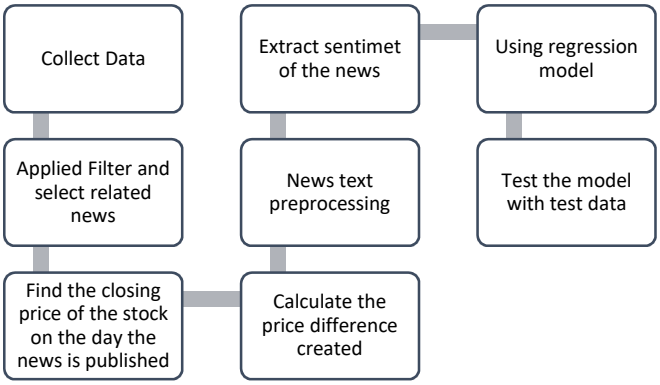

Figure 1 proposed model

D. Evaluation methodology

Root Mean Square Error (RMSE): RMSE is a popular measure of the prediction accuracy of a forecasting model [7]. It has a very intuitive interpretation in terms of relative error, represented mathematically as [9]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \quad (1)$$

where X_{obs} is observed values and X_{model} is modelled values at time/place *i.*

Also Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight [9].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |x_i - x| \quad (2)$$

A period of one day has been used to predict prices. This paper examines the relationship between news and the impact on stock prices. To show this relationship, we applied the sentiment analysis data and the price difference between the day before the news was published and the day of the news.

IV. TEST RESULTS

We have used several regression models to predict the changes that occurred after the news was published. In these experiments, the data of positive and negative sentiments were used separately and also positive and negative sentiments were used in combination, as it can be seen, the results were not significantly different and therefore it can be concluded that only withusing positive or negative news, we can predict stock price changes.

We repeated the experiments with random data. As expected, using sentiments data to predict changes leads to much better and more acceptable results.

Manual tagging is slow and expensive. In our experiments, it was shown that using a computer to tag, results is similar to manually tagging, and so for online price change prediction systems, computer tagging can be used with confidence. As can be seen from the output tables, the results are not significantly different, so it can be concluded that using TextBlob alone is sufficient and there is no need for manual tagging.

Table 1 RMSE error using positive sentiments

Regression Model	RMSE_test	RMSE_train
SVR_rbf	0.39	0.38
SVR_linear	0.36	0.35
SVR_poly	0.35	0.35
RegressionTree	0.47	0.47
RandomForrest	0.46	0.45
BayesianRidge	0.35	0.35
LASSO	0.34	0.33

Table 2 MAE error using positive sentiments

Regression Model	MAE_test	MAE_train
SVR_rbf	0.26	0.23
SVR_linear	0.26	0.27
SVR_poly	0.29	0.27
RegressionTree	0.17	0.15
RandomForrest	0.17	0.16
BayesianRidge	0.29	0.27

LASSO	0.31	0.28

Table 3 RMSE error using negative sentiments

Regression Model	RMSE_test	RMSE_train
SVR_rbf	0.32	0.32
SVR_linear	0.31	0.32
SVR_poly	0.31	0.31
RegressionTree	0.41	0.40
RandomForrest	0.40	0.40
BayesianRidge	0.31	0.41
LASSO	0.32	0.41

Table 4 MAE error using negative sentiments

Regression Model	MAE_test	MAE_train
SVR_rbf	0.23	0.23
SVR_linear	0.26	0.25
SVR_poly	0.25	0.25
RegressionTree	0.16	0.12
RandomForrest	0.12	0.15
BayesianRidge	0.26	0.25
LASSO	0.26	0.25

Table5 RMSE error using positive and negative sentiment

Regression Model	RMSE_test	RMSE_train
SVR_rbf	0.35	0.35
SVR_linear	0.36	0.37
SVR_poly	0.35	0.34
RegressionTree	0.46	0.44
RandomForrest	0.47	0.45
BayesianRidge	0.45	0.35
LASSO	0.34	0.34

Table 6 MAE error using positive and negative sentiment

Regression Model	MAE_test	MAE_train
SVR_rbf	0.25	0.23
SVR_linear	0.27	0.27
SVR_poly	0.29	0.28
RegressionTree	0.17	0.14
RandomForrest	0.16	0.16
BayesianRidge	0.27	0.25
LASSO	0.28	0.27

Table 7 RMSE error Using random data tagging Vs. sentiment analysis data

Regression Model	Random Data tagging	Sentiment analysis data
SVR_rbf	0.62	0.35
SVR_linear	0.50	0.36
SVR_poly	0.55	0.35
RegressionTree	0.53	0.46
RandomForrest	0.51	0.47
BayesianRidge	0.48	0.45
LASSO	0.44	0.34

Table 8 RMSE error, Manual tagging Vs. Computer tagging

Regression Model	Manual tagging	Computer tagging
SVR_rbf	0.39	0.35
SVR_linear	0.35	0.36
SVR_poly	0.40	0.35
RegressionTree	0.53	0.46
RandomForrest	0.51	0.47
BayesianRidge	0.37	0.45
LASSO	0.37	0.34

V. Conclusion

Nowadays, the stock market has become an important channel for raising funds for investors. The stock market is considered as a measure of economic and financial activities in a country or region. Predicting stock price movements is one of the most challenging tasks in the financial world. Because the stock market is dynamic, complex, nonlinear and non-parametric in nature, accurate forecasting of stock price variable is critical to developing a trading strategy. One of the main sources that investors consider is news, which they decide to invest based on the news that is published from the stock Therefore, we can use the news and the amount of positive or negative sentiments it evokes in the reader to predict changes.

In experiments, we found that using sentiment analysis data works better than random data. Also, since manual tagging is expensive and slow, the study showed that the results of both methods will be almost the same.

In addition, we showed that using one of the two positive or negative sentiments is enough to predict stock price changes and it is not necessary to have data on both sentiments.

References

[1] D. Duong, T. Nguyen, and M. Dang, “Stock market prediction using financial news articles on ho chi minh stock exchange,” in Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication

[2] Y. Wang, D. Seyler, S. K. K. Santu, and C. Zhai, “A study of feature construction for text-based forecasting of time series variables,” in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.

[3] Tianyi Wang , Ke Lu , Kam Pui Chow , Qing Zhu (2020). "COVID-19 sensing: negative sentiment analysis on social media in China via BERT

[4] Khedr, A. E. and N. Yaseen (2017). "Predicting stock market behavior using data mining technique and news sentiment analysis." International Journal of Intelligent Systems and Applications 9(7): 22.

[5] Joshi Kalyani, Prof. H. N. Bharathi, Prof. Rao Jyothi (2016). "Stock trend prediction using news sentiment analysis." arXiv preprint arXiv:1.607.01958

[6] Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan (2015). Deep learning for event-driven stock prediction. Twenty-fourth international joint conference on artificial intelligence.

[7] Ananthi, M. and Vijayakumar, K., 2021. Stock market analysis using candlestick regression and market trend prediction (CKRM). Journal of Ambient Intelligence and Humanized Computing, 12(5), pp.4819-4826.

[8] Javed Awan, M., Mohd Rahim, M.S., Nobanee, H., Munawar, A., Yasin, A. and Zain, A.M., 2021. Social media and stock market prediction: a big data approach. MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua, 67(2), pp.2569-2583.

[9] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better