IJICTR
International Journal of Information & Communication Technology Research

Volume 16- Number 3 – 2024 (54 -62)

Investigating Sentiment Analysis of News in Stock Market Prediction Utilizing Machine Learning Techniques

Golshid Ranjbaran
Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University,
Tehran, Iran
Golshid.ranjbaran@srbiau.ac.ir

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

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

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

Received: 15 October 2023 – Revised: 10 January 2024 - Accepted: 2 April 2024

Abstract—**In the stock market, which is a dynamic, complex, nonlinear and non-parametric environment, accurate prediction is crucial for trading strategy. It is assumed that news articles affect the stock market. We investigated the relationship[1] between headline's sentiment of news and their impact on stock prices changes. To show this relationship, we applied the sentiment data and the price difference between the day before the news was published and the day of the news, to machine learning regression and classification models. Regression is used to predict changes and classification is used to decide whether to buy or sell stocks. We used three stock datasets named Apple, Amazon and AXP and the results are shown in the mentioned dataset that using news with negative sentiments can make predictions just as correctly as using news with both positive and negative sentiments. In regression and classification models, Random Forest outperformed other machine learning algorithms in predicting stock price changes using news sentiment analysis. Additionally, we depicted that the results of computer and human tagging were almost similar, showing that using computer tools for text tagging will allow to tag text much more quickly and easily.**

Keywords: News, stock price prediction, machine learning, Regression model, Classification model

Article type: Research Article

Publisher: ICT Research Institute

* Corresponding Author

I. INTRODUCTION

In finance studies, 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 huge amount of data every day, it is very difficult to consider all the current and past information to predict the future trend of stocks. There are basically two analyses 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[1,2].

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 decide 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 a share of Amazon, and thus a positive news can lead to growth the price [3-5].

Stock market predicting 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 are 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 evokes in the reader can be considered as a very important factor in predicting the stock price [6-8].

Both regression and classification methods are used in stock market predictions. In the regression method, past data is used to predict the amount of changes and according to the size of the changes, it can be decided to buy and sell stocks. The classification method uses past data for classification. Based on the amount of change, our data is categorized into three sentiment categories: positive, zero, and negative. Therefore, we predicted changes utilizing sentiments. Related works are reviewed below.

In this paper, we investigate sentiment analysis of news to predict stock market prices. The study focuses on analyzing positive and negative sentiments and proposes a new method to explore the relationship between stock price changes and news sentiment. Also, as a new finding, we demonstrate the advantages of automated sentiment analysis tools through a comparative analysis between manual and automatic sentiment tagging, thereby enhancing the field's understanding of sentiment analysis in stock market prediction.

The paper focuses on examining the relationship between news sentiment in headlines and stock price changes. Machine learning models using regression and classification techniques are employed to predict stock price fluctuations. The study investigates correlation rather than causality, as there is no causal relationship between news sentiment in headlines and stock price changes. Analyzing sentiment data and comparing it with price differences between the day before news publication and the day of news release demonstrates the correlation between these two variables.

The rest of this paper is organized as follows. Section 2 includes the related works. In Section 3 the explanation and reference of the data are provided, In addition, in the mentioned Section the research methodology and evaluation metrics are stated. Section 4 presents the results of experiments, and finally Section 5 presents the conclusions of the article. Furthermore, references can be found in Section 6.

II. RELATED WORKS

Predicting stock price trends is an attractive area of research because more accurate predicts 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 [9,10]. In light of this, price prediction can benefit from sentimental news extraction. Reviewing articles in an evolutionary manner based on their historical occurrence is a fundamental aspect of conducting a comprehensive literature survey.

Authors in [19] aims to assess the effectiveness of using the Chicago Board Options Exchange Market Volatility Index (VIX) in combination with Support Vector Machines (SVMs) to predict changes in the S&P 500 index on a weekly basis. The study analyzed data from January 2000 to December 2011 and implemented a trading simulation to evaluate both statistical efficiency and economic performance. The inputs considered included traditional technical trading rules and indicators such as the Relative Strength Index, Moving Average Convergence Divergence, VIX, and daily returns of the S&P 500. The SVM model identified optimal buying and selling opportunities in the market. Results revealed that SVM using VIX produced superior outcomes compared to the Buy and Hold strategy or SVM models without VIX. The inclusion of VIX proved particularly influential during bearish periods, allowing for a reduction in Maximum Drawdown and annualized standard deviation.

Work done 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". They proved with experiments that news headlines should be sufficient for textual features and thus improve the share price prediction [14].

In [15], the FA (Firefly Algorithm) was modified by introducing a dynamic adjustment strategy and an opposition-based chaotic strategy. In that context, the Modified FA (MFA) was combined with an SVR to propose a novel hybrid forecasting model (SVR-MFA) for stock prices. The forecasting results with the SVR-MFA were more robust and moreover outperform the results of the SVR-CFA (Support vector regression with Chaotic Firefly Algorithm) [16] as well as the other optimization techniques PSO-GA (Particle Swarm Optimization - Genetic Algorithm) applied with SVR models for the prediction of six variables in the Shanghai stock market.

The goal of the paper in [12] 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: KNN (K-Nearest Neighbor) and simple Bayesian algorithm. Khedr et al. separates the model into two stages, the first is to determine the positive or negative polarity of the news using a simple Bayesian algorithm. In the second stage, the output of the first stage is entered as an input, along with the past stock price, into the K-NN algorithm and combined algorithms.

Kalyani et al. 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 showed that Random Forest and SVM (Support Vector Machine) worked well in all experiments [13].

Sedighi in [17] proposed a novel upgraded prediction model deploying an ABC (Artificial Bee Colony) algorithm, an ANFIS (Adaptive Neuro-Fuzzy Inference System), and SVM. In the proposed ABC-ANFIS-SVM approach, the ABC, ANFIS and SVM algorithms was used, respectively, to optimize the technical indicators, to forecast long-run stock price fluctuations, and to establish a link between the technical indicators and the stock price, as well as to further improve the forecasting accuracy of the model. The results obtained with the proposed method were more precise in the prediction of the US stock market than those from the 20 other forecasting models. The discussed studies on SVM-based approaches are summarized with their key topics in Table 1. It is noteworthy that, in this figure, we have not included the studies [18,19] that only utilized the SVM method without merging it with at least one other soft computing technique for making stock market predictions [20].

Usmani [4] proposed a new stock prediction model called WCN-LSTM(Weighted and Categorized News with Long Short-Term Memory) that incorporates news categories weighted according to their relevance with the target stock. The model utilizes a LSTM neural network and features such as hybrid input, sentiment analysis, and deep learning. Experiments were conducted on the Pakistan Stock Exchange using different sentiment dictionaries and time steps. The results show that WCN-LSTM outperforms the baseline model and other existing prediction models. The sentiment lexicon HIV4 and time steps 3 and 7 optimize the prediction accuracy.
Banerjee and er.el in [10] aims to develop a predictive model that classifies news headlines about Indian companies as either impactful or not impactful, based on past reactions to similar news. The experimental results highlight that the performance of classification algorithms depends on the chosen feature extraction technique, and vice-versa. Additionally, the study demonstrates that machine learning models combined with appropriate feature extraction techniques can reasonably predict market reactions to published news.

TABLE I. OVERVIEW OF RELATED WORKS

Ref	Main Approach	Dataset	Type of Input Data	Performance index
[4]	WCN-LSTM	Pakistan Stock Exchange (PSX)	financial news, stock market trends	accuracy
[15]	SVM/SVR	6 securities from SSE	Closing prices	RMSE
[17]	SVM/SVR	DOW 30, NASDAQ 100, S&P 500	Technical indicators	RMSE
[21]	Decision tree	Enron stock	Closing prices	accuracy
[22]	Classifier ensembles	KOSPI 200	Technical indicators	Accuracy, precision, recall, F1 score

[23]	(SVM, RF, ANN) + feature selection	Chinese A-share market	Technical, economic, and fundamental variables	Accuracy, sharp ratio, annualized return

III. RESEARCH METHODOLOGY

In this section, the research methodology of the paper is explained in five parts. Firstly, the dataset used is described. Secondly, the preprocessing steps undertaken are outlined. Thirdly, the predictions made are discussed. Fourthly, the proposed method is presented. Finally, the evaluation methodology used to assess the performance of the proposed method is explained.

A. Dataset

The dataset used in our work includes 1048575 news that has been collected over a period of eighteen years. This dataset is available to the public through this link[1]. It is a combination of all the news that is happening around the world. In this work we only needed news that was somehow related to the stock under review, thus, by applying a filter, we have selected only those containing a keyword of the stock in their headline. For example, to select news related to Apple company keywords such as 'apple', 'aapl[2]' are applied to the main data and the related news was extracted. In Table 2, three datasets extracted from the original dataset are presented, along with the number of positive and negative news articles for each stock.

TABLE II. DATASET SPECIFICATIONS

Stock name	Total Number of news	Positive news	Negative news
AXP[3]	362	225	137
Apple	1045	680	365
Amazon	298	157	141

B. Preprocessing

The mentioned data only includes the news that has been collected from different news agencies and the stock price was received separately from yahoofinance.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 models. In this stage of preprocessing, first all numbers, symbols, abbreviations, ineffective words, etc. were removed, and then TextBlob[4] was used as a tool to extract the sentiment of each news. This tool gives each news a number between 0 and 1 which is the sentiment of that news.

C. Prediction models

In this article, there are two sort of models that were used to predict changes in stock prices. Models for regression and classification. Regression analysis is a group of statistical procedures used in statistical modeling to calculate the relationships between a dependent variable and one or more independent variables. Regression analysis is frequently utilized in the fields of prediction and forecasting, where it has significant overlap with machine learning.

In machine learning, the term "classification" refers to a challenge involving predictive modeling in which a class label is predicted for a specific sample of input data. A model will determine the optimal way to map samples of input data to particular class labels using the training dataset.

In this article, Decision Tree, Random Forest, Lasso, Support Vector Regression (SVR) and Baysian Ridge are some of the regression models that are employed. Additionally, Decision Trees, Random Forests, K-NN, Support Vector Classifier (SVC) and Naïve Bayes are among the models utilized for classification.

D. Proposed method

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

- ○ Does the news have a meaningful impact on the stock market?
- ○ Positive or negative news: which one leads to more precise predictions?
- ○ How much more accurate is manual sentiment tagging at making predictions compared to computer sentiment tagging?

Prices have been predicted over a one-day timeframe. This paper investigates the relationship between news and its effect on stock prices. We used the sentiment analysis data and the price changes between the day before the news was announced and the day of the news release to demonstrate this relationship.

Our model for predicting changes in stock prices using news and sentiment are shown in Fig. 1 and 2. Firstly in Fig. 1, the stock-related news are extracted. Then the stock price changes triggered by this news are calculated. It has been measured how much the news has generated both positive and negative sentiment. Stock price changes have been predicted using these sentiments. The regression models learn and predict price changes using news sentiments.

Secondly in Fig. 2, the derived price difference was then split into positive and negative classes to predict the ‘price class’ for the next day. In this manner, the class is positive if the difference is positive, and the class is negative if the difference is negative. Finally, classification models were developed using these data. As a result, the stock class for each day is predicted using the sentiment obtained from the news.

The data were split into 80% for training and 20% for testing. We used a k-fold cross-validation with k=5. The results reported below are the results obtained on the test data set. The simulation is implemented in the

[1] https://www.kaggle.com/therohk/million-headlines

[2] The symbol of Apple company in the stock market

[3] American Express Company

[4] TextBlob is a Python library for processing textual data

Python programming environment using the scikit-learn library. Therefore, we have examined several algorithms to predict the price change that has occurred. The results are shown later.

In the analysis that has been done, 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 automatic tagging'. Indeed, manual tagging is costly, and much more news can be tagged using automatic tagging. This tagging can be done online and quickly. Therefore, it is possible to act very quickly and in real time to predict changes of stocks.

Additionally, in order to properly respond to the question: “whether the news actually affects the stock market or not”, we compared the prediction models with news sentiment data and random data. Random data is the data that is generated randomly and sentiment data are output of the news from the TextBlob library.

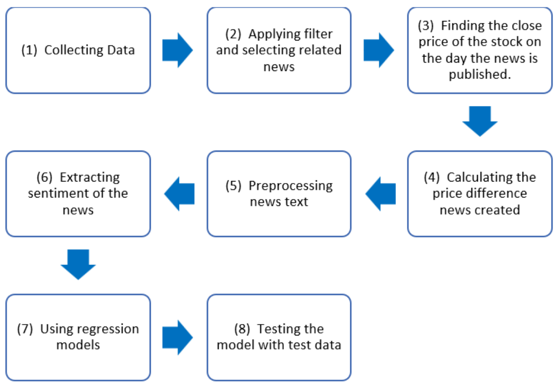
(1) Collecting Data
(2) Applying filter and selecting related news
(3) Finding the close price of the stock on the day the news is published.
(4) Calculating the price difference news created
(5) Preprocessing news text
(6) Extracting sentiment of the news
(7) Using regression models
(8) Testing the model with test data

Figure 1. Proposed model in Regression

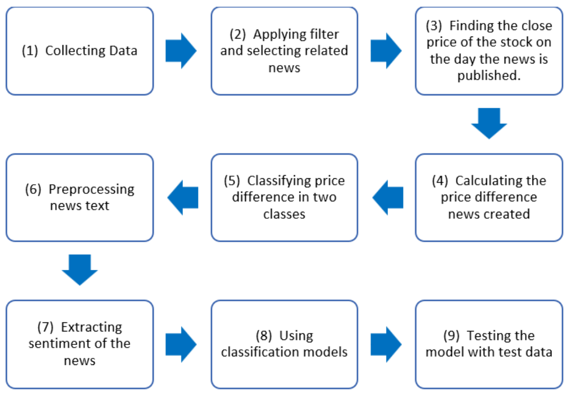
(1) Collecting Data
(2) Applying filter and selecting related news
(3) Finding the close price of the stock on the day the news is published.
(4) Calculating the price difference news created
(5) Classifying price difference in two classes
(6) Preprocessing news text
(7) Extracting sentiment of the news
(8) Using classification models
(9) Testing the model with test data

Figure 2. Proposed model in classification

E. Evaluation methodology

Various criteria have been used to evaluate the proposed model. RMSE (Root Mean Square Error), RRMSE (Relative Root Mean Squared Error) and MAE (Mean Absolute Error) can be mentioned among the criteria examined in regression models. In addition,, the accuracy and f1_Score criterion has been used to evaluate and compare the classification models. In Eq (1), the RMSE is shown to measure the prediction accuracy of a predicting model [15]. It has a very intuitive interpretation in terms of relative error, represented mathematically as [17]:

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

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

Moreover, Mean Absolute Error (MAE) in Eq (2) measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - X| \qquad (2)$$

By calculating the harmonic mean of a classifier's precision and recall, the F1-score in Eq(3) for classification combines these two metrics into a single one. It is primarily used to compare the performance of two classifiers. It should be noted that the ability of a classification model to return only the data points in a class is known as precision Eq (4).

$$F1_{Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (3)$$

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives+FalsePositives}} \qquad (4)$$

Accuracy in Equation (5) is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

$$Accuracy = \frac{TruePositives+TrueNegatives}{All\ Samples} \qquad (5)$$

Furthermore, the statistic known as "Sensitivity" in Eq (6) measures a model's potential to predict true positives for each available class.

$$Sensitivity = \frac{TruePositives}{TruePositive+FalseNegative} \qquad (6)$$

Another metric used in the evaluation of classification algorithms is known as” Specificity”. This criterion (Eq (7)) measures a model's potential to predict true negatives for each available class. Specificity determines a model’s ability to predict if an observation does not belong to a specific class.

$$Specificity = \frac{TrueNegatives}{TrueNegative+FalsePositive} \qquad (7)$$

IV. Experimental Results

To predict the price changes that occurred for Apple, Amazon, and AXP companies upon the news release, we utilized a variety of regression models. As can be seen in Fig. 3, positive and negative sentiments were examined independently in these experiments, as well as positive and negative sentiments in combination. In order to conduct this experiment, we aggregated datasets from three stocks and subsequently partitioned the data based on the news classification, specifically into positive, negative, and a combination of both categories.

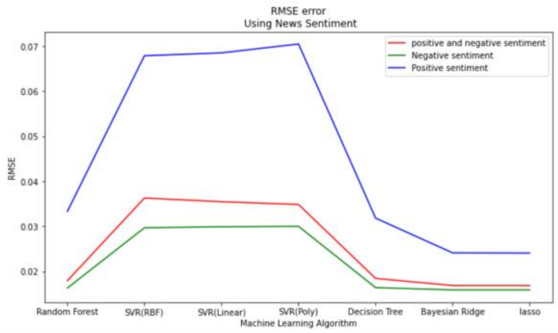

Figure 3. RMSE error, Using News Sentiment

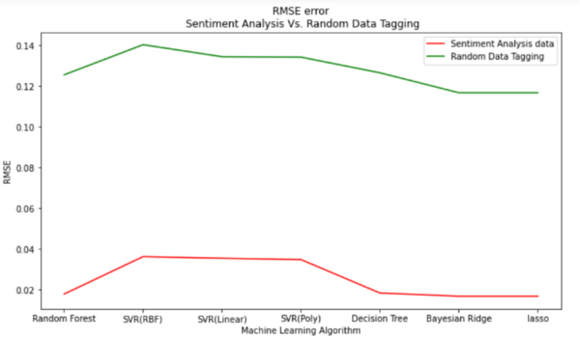

Figure 4. RMSE error, Using random data tagging Vs. sentiment analysis data

From and overall perspective, we proposed a method that analyzes news sentiment using the TextBlob library, investigated the correlation between news sentiment and stock price changes and employed regression and classification models to predict stock price fluctuations based on sentiment analysis. In addition, we compared manual and automatic tagging methods while examining the impact of news on the stock market.

To demonstrate how news affects predicts, an experiment was designed. Stock data tags were assigned randomly. With this random data, regression prediction models were trained to predict stock price changes. The outcomes were then compared with a sentiment analysis of the news, as shown in Fig. 4. In comparison to using random data tagging, the error that was previously derived via sentiment analysis of the news is significantly lower.

Another test was designed to compare the error rate of manual tagging with tagging using Python modules (textblob), which in this paper we call computer tagging. The data was first manually tagged, then the generated data was utilized to train regression models, and lastly the outcomes were assessed against the sentiment analysis performed with Python (Fig. 5). The findings indicate that there is no discernible difference between these two kinds of tagging models and that it is reliable to use the tools provided for data tagging. This drastically cuts down on the time and expense of making predictions.

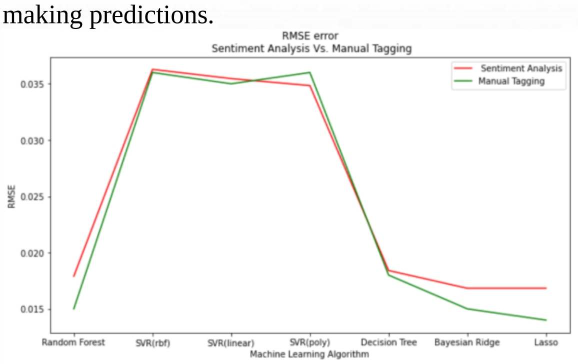

Figure 5. RMSE error, Manual tagging Vs. Computer tagging

Finally, Table 3 presents a detailed analysis and comparison of the efficiency and error of the prediction model for the three stocks of Apple, AXP, and Amazon. Initially, regression modeling was employed for prediction. It took the sentiments obtained from news as an input and stock price changes as an output. To investigate the relation between news and price changes, five machine learning algorithms were applied: SVR, Random Forest, Decision Tree, Bayesian Ridge, and Lasso. It should be mentioned that three distinct kernels have been used to test the SVR. Moreover, as mentioned in the proposed model section, positive and negative price changes that affected the stock were classified into two classes. The classification's output is the class of stock price changes, while the classification's input is sentiment analysis performed on news headlines. In this experiment, five classifications named SVC, Random Forest, Decision Tree, KNN, and Naive Bayes have been used. In the following, Table 2 is summarized in figures 6 and 7, providing the reader with a better overview of the obtained results.

TABLE III. RESULTS OF REGRESSION AND CLASSIFICATION MODELS ON APPLE, AXP, AMAZON STOCK

APPLE stock							
Regression							
Evaluation parameters	SVR _rbf	SVR_ linear	SVR _poly	Regression tree	Random forest	Bayesian Ridge	Lasso
RMSE	0.06	0.06	0.05	0.04	0.04	0.04	0.04
MAE	0.04	0.04	0.04	0.01	0.01	0.01	0.01
Classification							
	SVC _rbf	SVC_ linear	SVC _poly	Decision tree	Random forest	KNN	Naïve bayes
F1_score	0.77	0.77	0.76	0.78	0.80	0.76	0.60
Accuracy	0.96	0.96	0.95	0.96	0.96	0.96	0.94
Sensitivity	0.75	0.74	0.73	0.73	0.75	0.73	0.65
Specificity	0.98	0.98	0.98	0.98	0.98	0.98	0.96
Precision	0.97	0.97	0.97	0.97	0.97	0.97	0.94
AXP stock							
Regression							
Evaluation parameters	SVR _rbf	SVR_ linear	SVR _poly	Regression tree	Random forest	Bayesian Ridge	Lasso
RMSE	0.10	0.10	0.10	0.10	0.10	0.09	0.09
MAE	0.07	0.07	0.07	0.01	0.06	0.05	0.05

Classification							
	SVC _rbf	SVC_ linear	SVC _poly	Decision tree	Random forest	KNN	Naïve bayes
F1_score	0.55	0.52	0.62	0.70	0.77	0.62	0.60
Accuracy	0.95	0.94	0.96	0.96	0.97	0.96	0.94
Sensitivity	0.45	0.45	0.62	0.60	0.70	0.62	0.58
Specificity	0.98	0.98	0.98	0.99	0.99	0.98	0.97
Precision	0.95	0.95	0.96	0.98	0.98	0.96	0.95
AMAZON stock							
Regression							
Evaluation parameters	SVR _rbf	SVR_ linear	SVR _poly	Regression tree	Random forest	Bayesian Ridge	Lasso
RMSE	0.11	0.11	0.12	0.10	0.10	0.09	0.09
MAE	0.09	0.08	0.09	0.07	0.07	0.07	0.07
Classification							
	SVC _rbf	SVC_ linear	SVC _poly	Decision tree	Random forest	KNN	Naïve bayes
F1_score	0.57	0.53	0.55	0.70	0.76	0.62	0.58
Accuracy	0.92	0.91	0.90	0.93	0.95	0.91	0.90
Sensitivity	0.66	0.66	0.71	0.85	0.83	0.71	0.62
Specificity	0.94	0.93	0.92	0.94	0.97	0.94	0.93
Precision	0.91	0.90	0.89	0.93	0.96	0.92	0.89

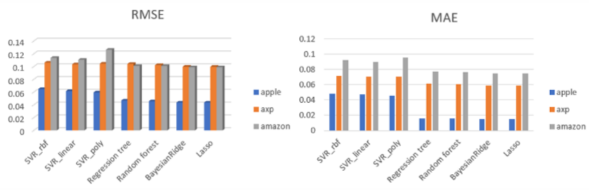

Figure 6. Bar chart of regression errors

V. Discussion and Results

As can be shown in Fig .3, the outcomes of negative sentiments and the combination of both sentiments have extremely strong predictive accuracy and are not significantly different. In regression prediction models, it was the positive sentiments that had the highest errors. As a result, we can estimate changes in stock price with a level of accuracy that is acceptable by having a negative or combination of negative sentiments. The results demonstrate that the stock prediction accuracy using negative sentiments is similar to that of a combination of positive and negative sentiments and both are remarkably high.
In terms of regression prediction models, it was observed that the highest errors were associated with positive sentiments. Consequently, by incorporating negative sentiments or a combination of positive and negative sentiments, we can reasonably estimate fluctuations in stock prices with an acceptable level of accuracy.

In Figure 4, we conducted a comparative analysis between randomly tagging and sentiment tagging approaches. Through experimental evaluation and regression modeling, we measured the RMSE as an indicator of performance. The results clearly demonstrate that sentiment tagging outperforms random tagging by a significant margin.

In addition, Fig. 5 illustrates the utilization of computer-based tagging in comparison to manual tagging. The objective of this analysis is to assess the degree of comparability between the two approaches. Our findings demonstrate that computer tagging yields comparable results to manual tagging, suggesting its effectiveness and potential as a viable alternative in the tagging process.

In Table. 3 based on the results presented in the paper, the f1_score and Precision values provide important insights. The high Precision value indicates that the algorithm has successfully identified a significant number of positive samples, which is a valuable outcome. It is worth noting that the stock Apple exhibited higher Precision compared to stocks AXP and Amazon, due to its larger dataset size. Amongst all implemented models, the highest F1 scores are achieved by the Random Forest algorithm.
In addition, the decrease in f1_score relative to Precision suggests a potential data sample imbalance between positive and negative news. To mitigate this issue, incorporating data augmentation techniques may help increasing the f1_score.

The findings indicate that the performance of different SVM kernels is close to each other and that SVM produces less appropriate results in predictions than other regression and classification models. It was determined that random trees performed extremely well at predicting price changes. The difference between the two evaluation metrics of sensitivity and specificity is caused by an unbalanced distribution of positive and negative samples in the classification. Sensitivity is the metric that evaluates a model's ability to predict true positives of each available class, and specificity is the metric that evaluates a model's ability to predict true negatives of each available class. And the distinction between these two metrics demonstrates unequivocally that our true positive and true negative sample sizes were different. In other words, price changes have been more negatively skewed than positively skewed. The considerable range of accuracy is a result of this issue.

The reader should be aware that all findings are expressed as percentages, and all data were standardized before being used in machine learning algorithms.

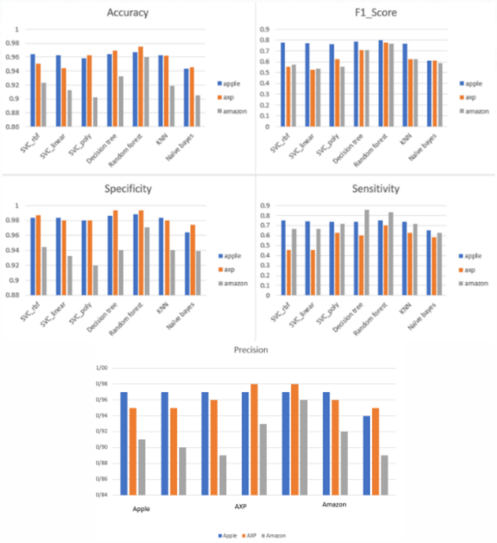

Figure 7. Bar chart of classification metrics

VI. Conclusion and Future Works

Nowadays, the stock market has become an important channel for raising funds for investors. Because the stock market is dynamic, complex, nonlinear, and non-parametric in nature, accurate forecasting of stock price variables is critical to developing a trading strategy. One of the main sources that investors consider is news, and 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 sentiment it evokes in the reader to predict changes.

In experiments that we had, it is clearly demonstrated that using sentiment analysis of news can be useful in anticipating changes in stock prices. The findings demonstrated that in the data examined in this research, while sentiment of negative news and the combination of positive and negative news sentiment had satisfactory and closely performance, positive news alone was unable to accurately predict stock changes. Analysis using several machine learning models including support vector regression (SVR) with different kernels, decision tree, random forest, k-nearest neighbors (KNN), and naive Bayes revealed that in the mentioned dataset, regardless of the kernel used, the support vector machine algorithm produced poor results. However, the random forest approach demonstrated acceptable results in both regression and classification prediction models. Sensitivity and specificity are separated by a significant distance. This matter declares that the proportion of positive and negative news was not equal; it is anticipated that in future research, the proportions of positive and negative news will be the same, resulting in more reliable results. The outcomes demonstrated that applying human tagging and applying computer tagging carried out with the use of machine learning techniques produced the same kinds of outcomes. This quick tagging can therefore be employed in real-time systems, drastically reducing the cost and time involved. As future work, based on the results and discussions presented in this paper, we suggest the utilization of larger and more balanced datasets to enhance the reliability and generalizability of the results, or employing data augmentation techniques to the current dataset to overcome the problems of size and unbalanceness.

References

[1] Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A., & Ganaie, I. A. (2023). Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. Procedia Computer Science, 218, 1067-1078.

[2] Biswas, S., Ghosh, S., Roy, S., Bose, R., & Soni, S. (2023). A Study of Stock Market Prediction through Sentiment Analysis. Mapana Journal of Sciences, 22(1).

[3] Duong D, Nguyen T, Dang M, editors. Stock market prediction using financial news articles on ho chi minh stock exchange. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication; 2016.

[4] Usmani, S., & Shamsi, J. A. (2023). LSTM based stock prediction using weighted and categorized financial news. Plos one, 18(3), e0282234.

[5] Wang Y, Seyler D, Santu SKK, Zhai C, editors. A study of feature construction for text-based forecasting of time series variables. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017.

[6] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv preprint arXiv:2304.07619.

[7] Puh, K., & Bagić Babac, M. (2023). Predicting stock market using natural language processing. American Journal of Business.

[8] Ranibaran, G., Moin, M. S., Alizadeh, S. H., & Koochari, A. (2021, December). Analyzing effect of news polarity on stock market prediction: a machine learning approach. In 2021 12th International Conference on Information and Knowledge Technology (IKT) (pp. 102-106). IEEE.

[9] Ashtiani, M. N., & Raahmei, B. (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. Expert Systems with Applications, 119509.

[10] Banerjee, P., Ananthakumar, U., & Singh, S. (2023). Predicting Impact of Published News Headlines Using Text Mining and Classification Techniques. In Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Proceedings of the ICNC-FSKD 2022 (pp. 983-992). Cham: Springer International Publishing.

[11] Wang T, Lu K, Chow KP, Zhu Q. COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model. Ieee Access. 2020;8:138162-9.

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

[13] Kalyani J, Bharathi P, Jyothi P. Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:160701958. 2016.

[14] Ding X, Zhang Y, Liu T, Duan J, editors. Deep learning for event-driven stock prediction. Twenty-fourth international joint conference on artificial intelligence; 2015.

[15] Zhang J, Teng Y-F, Chen W. Support vector regression with modified firefly algorithm for stock price forecasting. Applied Intelligence. 2019;49(5):1658-74.

[16] Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Applied soft computing. 2013;13(2):947-58.

[17] Sedighi M, Jahangirnia H, Gharakhani M, Farahani Fard S. A novel hybrid model for stock price forecasting based on metaheuristics and support vector machine. Data. 2019;4(2):75.

[18] Cao L. Support vector machines experts for time series forecasting. Neurocomputing. 2003;51:321-39.

[19] Rosillo R, Giner J, de la Fuente D. The effectiveness of the combined use of VIX and support vector machines on the prediction of S&P 500. Neural Computing and Applications. 2014;25(2):321-32.

[20] Kumbure MM, Lohrmann C, Luukka P, Porras J. Machine learning techniques and data for stock market forecasting: a literature review. Expert Systems with Applications. 2022:116659.

[21] Xuan Q, Zhou M, Zhang Z-Y, Fu C, Xiang Y, Wu Z, et al. Modern food foraging patterns: Geography and cuisine choices of restaurant patrons on yelp. IEEE Transactions on Computational Social Systems. 2018;5(2):508-17.

[22] Vickers NJ. Animal communication: when i'm calling you, will you answer too? Current biology. 2017;27(14):R713-R5.

[23] Xue X, Lu J. A compact brain storm algorithm for matching ontologies. Ieee Access. 2020;8:43898-907.

[24] Ananthi M, Vijayakumar K. Stock market analysis using candlestick regression and market trend prediction (CKRM). Journal of Ambient Intelligence and Humanized Computing. 2021;12(5):4819-26.

[25] Javed Awan M, Mohd Rahim MS, Nobanee H, Munawar A, Yasin A, Zain AM. 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. 2021;67(2):2569-83.

[26] Urquhart, B., Ghonima, M., Nguyen, D., Kurtz, B., Chow, C.W. and Kleissl, J., 2013. Sky-imaging systems for short-term forecasting. Solar energy forecasting and resource assessment, pp.195-232.

Golshid Ranjbaran received her M.Sc. degree in Artificial Intelligence from the Science and Research Branch University in 2017. Her interests lie in the fields of Machine Learning, Data Mining and Social Media Analysis.

Mohammad-Shahram Moin received his B.Sc. degree in Electronic Engineering from Amirkabir University of Technology in 1988, M.Sc. degree in Electronic Engineering from the University of Tehran's Faculty of Engineering in 1991, and Ph.D. degree in Electrical Engineering from Ecole Polytechnique of Montréal, Canada, in 2000. He is currently an Associate Professor at the ICT Research Institute, where he has led numerous projects in the fields of artificial intelligence, biometrics, multimedia, and big data. Dr. Moin has a teaching background in graduate courses such as pattern recognition, neural networks, data compression, data mining, digital signal processing, and stochastic processes. He has contributed to the publication of 45 journal papers, 7 book chapters, and 77 conference papers. His research interests include Artificial Intelligence, Pattern Recognition, Image Processing, Data Analysis and Biometrics. Dr. Moin is the head of the "Iranian Systems Scientific Society of Intelligent" and the Editor-in-Chief of the "Iranian Journal of Information and Communication Technology".

Sasan H. Alizadeh received his B.Sc. in Computer Hardware Engineering from Shiraz University, Shiraz, Iran. Also, he received his M.Sc. and the Ph.D. degrees in Computer Science from Amirkabir University of Technology, Tehran, Iran. He is a faculty member of the Department of Information Technology at ICT Research Institute (Iran Telecommunication Research Center), Tehran, Iran. His research interests include Statistical Machine Learning, Stochastic Processes, Bayesian Networks, Software Engineering, Social Network Analysis and Recommender Systems.

Abbas Koochari received his PhD in Computer Science with a focus on Artificial Intelligence in 2012. He is currently an Assistant Professor and a faculty member at the Science and Research Branch of Azad University. His research interests include image processing, machine vision, speech processing, natural languages, and deep learning.