Predicts stock price movements based on sentiment analysis of financial news

ABSTRACT

With the development of machine learning, predicting the pattern of products by analysing text has been actively investigated. Most of the existing studies only use the dataset of the target company, while some studies use the dataset of related companies in the Global Economy (GICS) field in the industry. However, we show that GICS has limitations in finding social media products due to the differences in the GICS market. To solve this limitation, we propose a method that shows the relationship and searches a group of workers with good relationships. Predicting stock prices using machine learning and multivariate learning techniques that combine information from target companies and their homogeneous groups. We conducted an experiment using one year of data and compared the results of the proposed method with existing methods. The results show that the proposed method is more predictive than existing methods in most cases. The results also show that the suitability of cluster analysis depends on the heterogeneity of sectors, and as the heterogeneity increases, cluster analysis of various groups is necessary.

Efficient Market Hypothesis is the popular theory about stock prediction. With its failure much research has been carried around prediction of stocks. This project is about taking non quantifiable data such as financial news articles about a company and predicting its future stock trend with news sentiment classification. If news articles have impact on stock market, this is an attempt to study relationship between news and stock trend. To show this, we created three different classification models which depict polarity of news articles being positive or negative. Observations show that RF and SVM perform well in all types of testing. Naïve Bayes gives good result but not compared to the other two. Experiments are conducted to evaluate various aspects of the proposed model and encouraging results are obtained in all of the experiments. The accuracy of the prediction model is more than 80% and in comparison, with news random labelling with 50% of accuracy; the model has increased the accuracy by 30%.

KEYWORDS

Text Mining - Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. You can use text mining to analyse vast collections of textual materials to capture key concepts, trends and hidden relationships.

Sentiment analysis - Sentiment analysis is the process of classifying whether a block of text is positive, negative, or neutral. The goal that Sentiment mining tries to gain is to be analysed people's opinions in a way that can help businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.,

Random Forest-Random Forest is a commonly used machine learning algorithm, trademarked by Leo Bierman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems.

SVM - Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

Stock trends - A trend is the overall direction of a market or an asset's price. In technical analysis, trends are identified by trendlines or price action that highlight when the price is making higher swing highs and higher swing lows for an uptrend, or lower swing lows and lower swing highs for a downtrend.

1. INTRODUCTION

In finance, the stock market and its structure are rare. Scientists are fascinated by the ability to catch the change and predict where it will lead next. Investors and market analysts study the market trends and plan their buying or selling strategies accordingly. Since the stock market generates so much data every day, it is difficult for someone to consider all the current and past data to predict the future of the stock. There are two ways to predict business trends One is technical analysis, the other is Fundamental analysis.

Technical analysis considers past price and volume to predict the future trend whereas Fundamental analysis. On the other hand, Fundamental analysis of a business involves analysing its financial data to get some insights. The efficacy of both technical and fundamental analysis is disputed by the efficient-market hypothesis which states that stock market prices are essentially unpredictable.

This research follows the Fundamental analysis technique to discover future trend of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative). If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down.

This research is an attempt to build a model that predicts news polarity which may affect changes in stock trends. In other words, check the impact of news articles on stock prices. We are using supervised machine learning as classification and other text mining techniques to check news polarity. And be able to classify unknown news, which is not used to build a classifier. Three different classification algorithms are implemented to check and improve classification accuracy. We have taken past one year data from Bajaj Finance as stock price and news articles.

2. LITERATURE SURVEY

Stock price trend prediction is an active research area, as more accurate predictions are directly related to more returns in stocks. Therefore, in recent years, significant efforts have been put into developing models that can predict for future trend of a specific stock or overall market. Most of the existing techniques make use of the technical indicators. Some of the researchers showed that there is a strong relationship between news article about a company and its stock prices fluctuations. Following is discussion on previous research on sentiment analysis of text data and different classification techniques. Nagar and Hahsler in their research [1] presented an automated text mining-based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analysed using Natural Language Processing (NLP) techniques. A sentiment metric, called News Sentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. They have used various open-source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of News Sentiment shows a very strong correlation with the actual stock price movement. Yu et al [2] present a text mining-based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified and then presented as a time series and compared with fluctuations in energy demand and prices. J. Bean [3] uses keyword tagging on Twitter feeds about airlines satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings. We have used the sentiment detection algorithm based on this research. This research paper [4] studies how the results of financial forecasting can be improved when news

articles with different levels of relevance to the target stock are used simultaneously. They used multiple kernels learning technique for partitioning the information which is extracted from different five categories of news articles based on sectors, sub-sectors, industries etc. News articles are divided into the five categories of relevance to a targeted stock, its sub industry, industry, group industry and sector while separate kernels are employed to analyse each one. The experimental results show that the simultaneous usage of five news categories improves the prediction performance in comparison with methods based on a lower number of news categories. The findings have shown that the highest prediction accuracy and return per trade were achieved for MKL when all five categories of news were utilized with two separate kernels of the polynomial and Gaussian types used for each news category.

3. RELATED WORK

The study conducted by Adebiyi et al. [6] evaluated the performance of an Artificial Neural Network (ANN) against an Autoregressive Integrated Moving Average (ARIMA) model using historical stock data of Dell Inc. on the New York Stock Exchange (NYSE). The research concluded that the ANN model performed slightly better than the ARIMA model and noted that incorporating macroeconomic and technical indicators could further improve the results.

In a 2019 study, Karmiani and colleagues [7] compared the performance of LSTM, Backpropagation, SVM, and Kalman filter for stock price prediction. They used historic data from Yahoo Finance for nine companies (Apple, Acer, Amazon, Google, HP, IBM, Intel, Microsoft, and Sony) and found that LSTM had the highest prediction accuracy and lowest variance among the models tested.

Chen et al. (2015) [8] used LSTM to predict stock prices in the China stock market, using historical data from the Shanghai and Shenzhen stock markets obtained from Yahoo finance as input features. They reported an accuracy of 27.2% and suggested that incorporating other features such as macroeconomic data and technical indicators would improve the model's performance.

In the study by Roondiwala et al. [10], LSTM was utilized to predict future stock prices of NIFTY50. Historical data, including high, low, open, and close prices, was obtained from the National Stock Exchange and used as input features. The RMSprop optimizer was employed with 500 epochs, resulting in a testing RMSE score of 0.00859. However, it is possible that normalized data were used to calculate the RMSE, rather than actual data. Additionally, the model's performance could have been improved by incorporating other factors, such as financial sentiments, that have a direct impact on stock prices.

In the study by Yu and Yan [11], data for six stock indices from various market environments were used, including the S&P 500, DJIA, N 225, HSI, CSI 300, and ChiNext index. In the first stage, the authors applied phase-space reconstruction (PSR), de-noising, and normalization to the data to improve the performance of the model. Four standard machine learning algorithms were compared: LSTM, MLP, SVR, and ARIMA. The results showed that the LSTM had the highest prediction accuracy among the algorithms compared.

In the work of Gao et al. the group applied four neural networks named Multilayer Perceptron (MLP), Long Short Term Memory (LSTM), Convolutional Neural Network (CNN) and one attention-based neural network — Uncertainty-aware Attention (UA)—to test the performance on predicting three stock market price: the SP500 index (most developed market), CSI300 index (less developed market) and Nikkei225 index (developing market) [12]. The results show that UA has the best performance among the alternative models. Furthermore, all models have better accuracy in the developed financial market than in developing ones.

In their study, Shahi et al. (2020) [13] investigated if incorporating financial news sentiments could improve the performance of stock price prediction using LSTM and GRU models. They used historical data from the Agricultural Development Bank Limited (ADBL) of Nepal and financial news headlines from ShareSansar Nepal,

from 20 March 2011 to 14 November 2019. The results showed that the performance of both LSTM and GRU models was significantly improved by including financial news sentiments as input features.

Kara et al.in 2011 [14] compared two classifiers —Artificial neural networks (ANN) and support vector machines (SVM) —to predict the direction of movement in the daily Istanbul Stock Exchange (ISE) National 100 Index.

The data from January 2, 1997 to December 3, 2007 were taken. Then technical indicators: Simple 10-day moving average, Weighted 10-day moving average, Momentum, Stochastic K%, Stochastic D%, Relative Strength Index (RSI), moving average convergence divergence (MACD), Larry William's R%, Accumulation/Distribution Oscillator, and Commodity Channel Index were selected as input variables. Experimental results showed that the average performance of ANN model (75.74%) was found significantly better than that of SVM model (71.52%).

In the work of Schoneburg, E. [15], the author analyzed the possibility of predicting stock prices on a short-term, day-to-day basis with the help of neural networks —Perceptron, Adaline, Madaline, and Backpropagation —by studying three important German stocks: BASF, COMMERZBANK, MERCEDES. The author achieved an

accuracy of up to 90% with a back propagation network. Moreover, he expresses that the selection of more suitable inputs for the network could improve the performance of the model.

K. Kohara et al. [16] used neural networks (NNs) for the prediction of the daily closing price of Tokyo stock price index (TOPIX) whether incorporating Event-knowledge (the daily headlines of the Japanese newspaper) could produce a better performance. The five inputs —Close: closing price of TOPIX, Exchange: the dollar-to-yen exchange rate (yen/dollar), Interest: an interest rate, Oil: the price of crude oil, and NY: New York Dow-Jones average of the closing prices of 30 industrial stocks —with and without Event-knowledge, feed to the NNs. The result shows that the performance of NNs is improved significantly by incorporating Event-knowledge.

Adebiyi A. A. et al. [17] used feed forward multilayer perceptron neural network with backpropagation whether incorporating fundamental analysis variables could produce a better performance than technical analysis variables only. The published stock data obtained from the Internet were used. The empirical results show that the performance of the model improved significantly by incorporating fundamental analysis variables for daily stock price prediction.

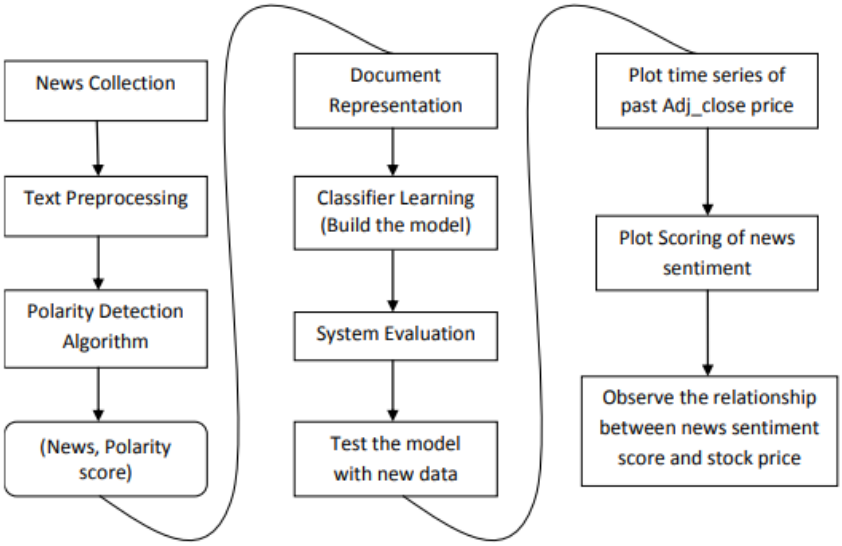
Selvin S. et al. [18] used Recurrent Neural Networks (RNN), Long Short Term Memory(LSTM), and Convolutional Neural Network (CNN) for short term stock price prediction using a sliding window approach. The window size was 100 minutes with 90 minute's information and prediction was made for the rest of the 10 minutes. Two companies from IT sector and one company from the Pharma sector of NIFTY were taken for the study. For their proposed methodology, CNN is identified as the best model.

During the COVID-19 pandemic in 2021, Binrong Wu. et al. [19] utilized the following machine learning models to forecast oil price, oil production, oil consumption, and oil inventory: CNN (Convolutional neural network), BPNN (Backpropagation neural networks), SVM (Support vector machines), LSTM (Long short-term memory), and RNN (Recurrent neural network). Their empirical findings suggest that information gleaned from social media platforms makes a major contribution to the process of forecasting oil prices, production levels, and consumption rates.

4. METHODOLOGY

System Design

Following system design is proposed in this project to classify news articles for generating stock trend signal.



This design can logically be seen as three phases with first column of blocks in phase 1, second column as phase 2 and third column contains blocks in phase 3.

- Result of phase 1 is news articles with its polarity score. This result is given as an input to the phase 2.
- In phase 2, text is converted in tied vector space so that it can be given to the classifier. Then three different classifiers are programmed for the same data to compare results. At the end of phase 2, we evaluate the results given by all classifiers and test for checking classifier performance for new news articles.
- In phase 3, we check for relationship between news articles and stock price data. We plot both the data using Python language and record the results. In the following sections, each block of the design is explained.

4.1 News Collection

We collected Bajaj Finance's data for past one year, from 1 Jan 2024 to 10 Sep 2024. This data includes major key events news articles of the company and daily stock prices of BAJFINANCE.NS for the same period. Daily stock prices contain six values as Open, High, Low, Close, Adjusted Close, and Volume. For integrity throughout the project, we considered Adjusted Close price as everyday stock price. We have collected this data from major news aggregators such as news.google.com, finance.yahoo.com.

Broadly speaking, this study uses two types of datasets — Historical Stock Data and financial news data. Stock market data consist of stock data between Jan 1, 2024, to Sep 10, 2024 accessed from Yahoo Finance [22], Google news. It consists of the open, close, high, low stock price as well as the shares traded (volume) on a particular day.

Overall description of Dataset

Dataset	Features	Source	Date	Frequency
Historical Stock Data	Open, Close, High, Low, Volume	• Yahoo • Google News	01-01-2024 to 10-09-2024	Daily
Financial News Data	News Headlines, News Body			At Least 5 news per day

4.2 Pre-Processing

Text data is unstructured data. So, we cannot provide raw test data to classifier as an input. Firstly, we need to tokenize the document into words to operate on word level. Text data contains more noisy words which are not contributing towards classification. So, we need to drop those words. In addition, text data may contain numbers, more white spaces, tabs, punctuation characters, stop words etc. We also need to clean data by removing all those words. For this purpose, we created own stop-word list which specifically contains stop words related to finance world and general English stop words. We built this using reference from [19]. This stop words list contains general words including Generic, names, Date and numbers, Geographic, Currencies. Also, to ignore words that appear in only one or two documents, we are considering minimum document frequency which considers words that appear in minimum three documents. Stemming is also important to reduce redundancy in words. Using stemming process, all the words are replaced by its original version of word. For example, the words 'developed', 'development', 'developing' are reduced to its stem word 'develop'. Some of the pre-processing is done before applying polarity detection algorithm. And some of them are applied after applying polarity detection algorithm.

4.3 Sentiment Detection Algorithm

For automatic sentiment detection of news articles, we use BeautifulSoap, Vader.

BeautifulSoap [23] is a Python library for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates our expectations.

VADER [24] is a pre-trained model that analyses people's opinions, sentiments, evaluations, attitudes, and emotions via computational treatment regarding polarity (positive/negative) and intensity (strength) in text. It relies on an English dictionary that maps lexical features to their semantic orientation as positive or negative.

Before inputting financial news obtained through web scraping into VEDAR for sentiment analysis, the data is first pre-processed. This involves removing any unnecessary text found in HTML tags and single or multiple blank spaces and escape sequences. However, the exclamation marks or question marks found in the news headline are not removed as they may add intensity and strength to the news. After preprocessing, the data is fed into VEDAR to determine its corresponding sentiment scores.

4.4 News Sentiment Score

Once the data are concatenated into a single frame based on the date column that exist in both datasets. The descriptive statistics of the combined data is illustrated to gain the initial insight of it. some metrics have large magnitude compared to others. To avoid features with large magnitude, dominate the feature with a small magnitude

The snapshot of the actual features used in the model

Date	Open	High	Low	Close	Adj Close	Volume	sentiment_score
01-01-2024	7336.95	7336.950195	7273	7299.0498	7262.59082	331489	0.9999
02-01-2024	7324	7445.100098	7280.049805	7430.0498	7392.93652	1112990	0.9995
03-01-2024	7444.95	7485.850098	7368.200195	7384.7998	7347.9126	553405	0.9983
04-01-2024	7560	7733.950195	7560	7705.5498	7667.06006	2911879	0.9999
05-01-2024	7734.95	7789	7673.100098	7711.1499	7672.63232	1052601	0.9999
08-01-2024	7677	7830	7631.299805	7736	7697.3584	1121209	0.9995
09-01-2024	7774	7810	7698	7725.4502	7686.86133	1247083	0.9999
10-01-2024	7716	7735.25	7660	7680.5498	7642.18506	559072	0.9999
11-01-2024	7680	7793.299805	7651.049805	7669.75	7631.43945	651741	0.9999
12-01-2024	7684	7718.950195	7617	7661.0498	7622.78271	808638	0.9995
15-01-2024	7697.95	7697.950195	7454	7478	7440.64697	997220	0.9996

Snapshot of the descriptive statistics of the features

	Date	Open	High	Low	Close	Adj Close	Volume	sentiment_score
count	141	141.000000	141.000000	141.000000	141.000000	141.000000	1.410000e+02	141.000000
mean	2024-04-16 03:14:02.553191424	6967.636185	7033.496485	6884.296082	6952.514884	6924.734375	1.291315e+06	0.983350
min	2024-01-01 00:00:00	6332.500000	6345.899902	6187.799805	6311.250000	6279.725098	3.314890e+05	-0.897900
25%	2024-02-21 00:00:00	6700.350098	6771.950195	6624.200195	6697.700195	6664.244629	8.249080e+05	0.999500
50%	2024-04-18 00:00:00	6937.850098	6995.000000	6845.000000	6910.100098	6881.006348	1.110184e+06	0.999600
75%	2024-06-10 00:00:00	7220.000000	7265.000000	7111.000000	7191.649902	7165.600098	1.408087e+06	0.999900
max	2024-07-31 00:00:00	7774.000000	7830.000000	7698.000000	7736.000000	7697.358398	6.818601e+06	0.999900
std	NaN	323.402162	324.088059	324.368287	322.880206	322.783761	8.844564e+05	0.162072

4.5 Document Representation

In order to reduce the complexity of text documents and make them easier to work with, the documents has to be transformed from the full text version to a document vector which describes the contents of the document. To represent text documents, we are using TF-IDF scheme. The higher tf-idf value a term gets, the more important it is. A high value is reached when the term frequency in the given document is high and when there are few other documents in the collection containing the given term/feature. This term weighting method tends therefore to filter out common terms by giving them a very low value.

4.6 Classifier Learning

As most of the research shows that SVM, Random Forest classification algorithms perform good in text classification.

Support Vector Machine (SVM) (Cortes & Vapnik, 1995) is a method for the classification of linear and nonlinear data and uses nonlinear mapping to transform the original training data into a higher dimension. Support Vector Machines then search for the linear optimal separating hyperplane, which essentially is a boundary which separates the records into classes. The Support Vector Machine finds the optimal hyperplane using support vectors (which can be characterised as the most significant training records) and margins (which are defined by the support vectors) (Han et al., 2011). The Support Vector Machine can be trained using several functions: the Linear Kernel Function, Quadratic, Gaussian Radial Basis (GRB), Multilayer Perceptron Kernel (MP) functions. Fig. 4 shows an example Support Vector Machine applied to ovarian cancer prediction using a linear separation of classes (Gaul et al., 2015). The most suitable training function is often experimentally selected. Support Vector Machines, when trained using the appropriate training function, are highly accurate and capable of modelling complex nonlinear decision boundaries (Han et al., 2011). They are also much less prone to overfitting than other computational intelligence methods and can be used for classification as well as prediction tasks. Support Vector Machines have been adopted to solve decision-making tasks on data which also include data obtained from images, microarray and other clinical data. So, we are considering these two algorithms to classify the text and check each algorithm's accuracy. We can compare all the results such as accuracy, precision, recall and other model evaluation methods. Two classification algorithms are implemented.

Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm,

the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

4.7 System Evaluation

We divided the data into train and test set. Also, we created unknown data set for classifier to check accuracy of classifier against new data. We evaluated all three classifiers performance by checking each one's accuracy, precision, recall, ROC curve area. The results are as given in the next section.

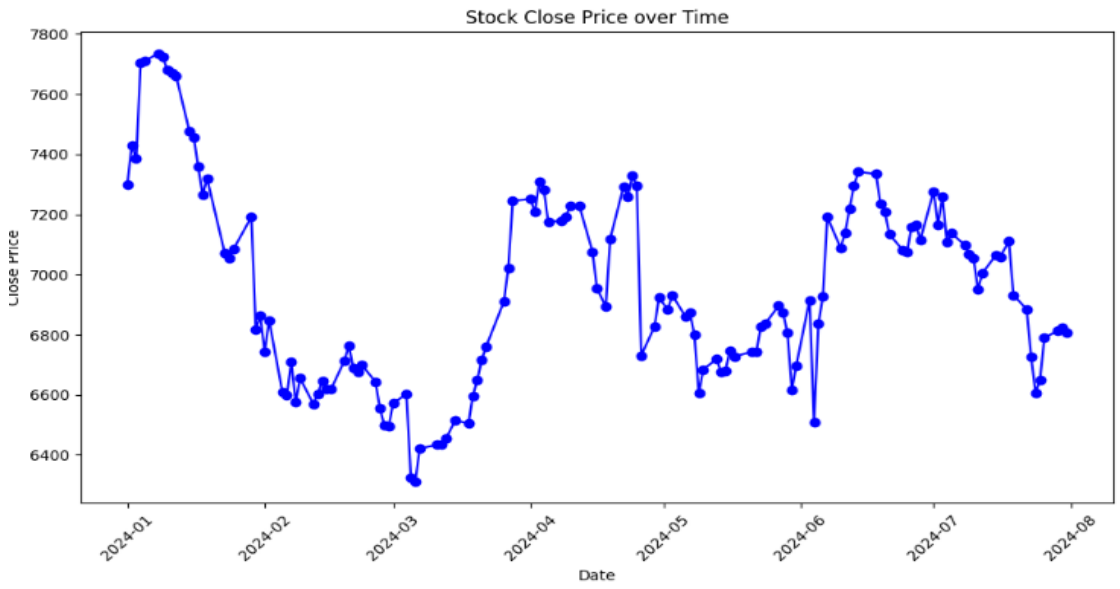
4.8 Evaluation

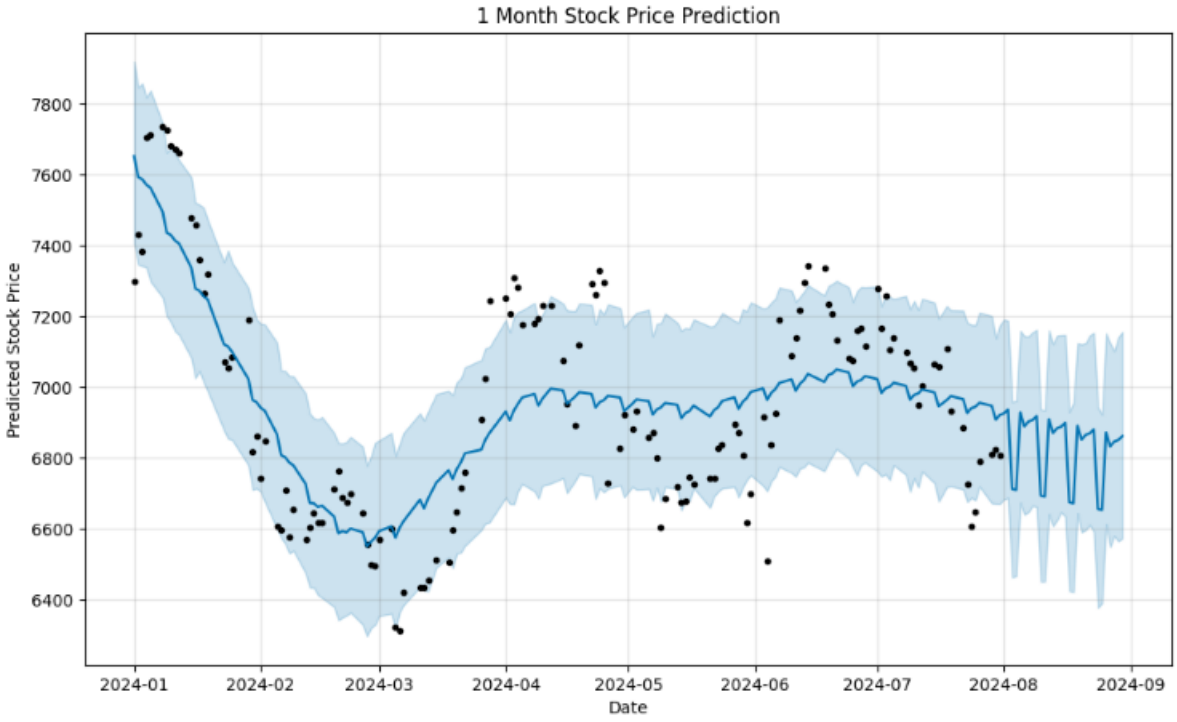
We tested the models using different testing options so that we can compare each method against different scenarios. Following are the test options on which we tested our models.

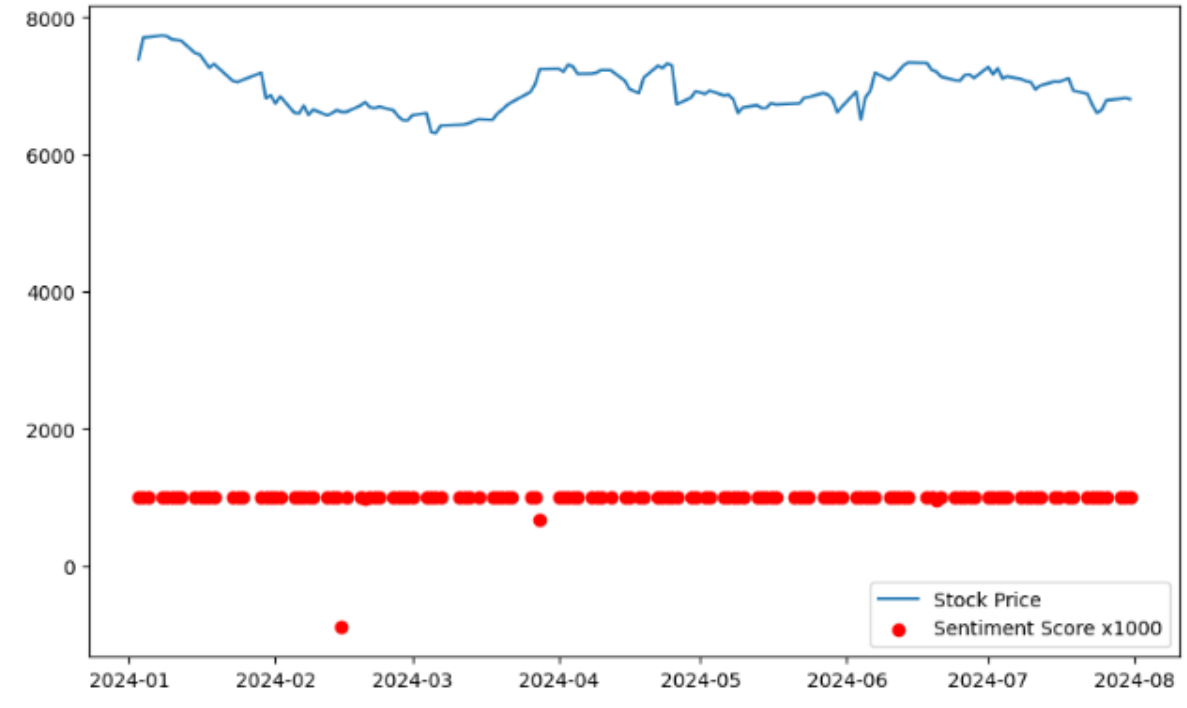
- 5 cross validations
- 10 cross validation
- 15 cross validation
- 70% Data split
- 80% Data split

	Test Option					
Classification Algorithm	**Classified**	**5 cross validations**	**10 cross validation**	**15 cross validation**	**70% Data split**	**80% Data split**
	Random Forest	79%	78%	76%	78%	76%
	SVM	74%	74%	73%	75%	74%
	ROC					
	Random Forest	.78	.77	.76	.78	.78
	SVM	.76	.75	.76	.78	.77
	Precision					
	Random Forest	.78	.78	.79	.76	.76
	SVM	.74	.76	.76	.74	.75
	Recall					
	Random Forest	.76	76	.75	.78	.76
	SVM	.74	.75	.74	.76	.76

4.9 Plot Time series




1 Month Stock Price Prediction
Predicted Stock Price
7800
7600
7400
7200
7000
6800
6600
6400
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
Date


8000
6000
4000
2000
0
Stock Price
Sentiment Score x1000
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08

5. CONCLUSION

Finding future trend for a stock is a crucial task because stock trends depend on number of factors. We assumed that news articles and stock price are related to each other. And news may have capacity to fluctuate stock trend. So, we thoroughly studied this relationship and concluded that stock trend can be predicted using news articles and previous price history. As news articles capture sentiment about the current market, we automate this sentiment detection and based on the words in the news articles, we can get an overall news polarity. If the news is positive, then we can state that this news impact is good in the market, so more chances of stock price go high. And if the news is negative, then it may impact the stock price to go down in trend. We used polarity detection algorithm for initially labelling news and making the train set. For this algorithm, dictionary-based approach was used. The dictionaries for positive and negative words are created using general and finance specific sentiment carrying words. Then preprocessing of text data was also a challenging task. We created own dictionary for stop words removal which also includes finance specific stop words. Based on this data, we implemented three classification models and tested under different test scenarios. Then after comparing their results, Random Forest worked very well for all test cases ranging from 88% to 92% accuracy. Accuracy followed by SVM is also considerable around 86%. Given any news article, it would be possible for the model to arrive on a polarity which would further predict the stock trend.

6. FUTURE WORK

I would like to extend this research by adding more company's data and check the prediction accuracy. For those companies where availability of financial news is a challenge, we would be using twitter data for similar analysis. We can also incorporate similar strategies for algorithmic trading.

7. REFERENCES

[1] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore.

[2] W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011, 5(3): 211-224.

[3] J. Bean, R by example: Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation, 2011.

[4] Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price Movements Based on Different Categories of News Articles, 2015 IEEE Symposium Series on Computational Intelligence.

[5] P. Hofmarcher, S. Theussl, and K. Hornik, Do Media Sentiments Reflect Economic Indices? Chinese Business Review. 2011, 10(7): 487-492.

[6] Adebiyi AA, Adewumi AO, Ayo CK. Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*. 2014;2014. doi: 10.1155/2014/614342.

[7] Karmiani D, Kazi R, Nambisan A, Shah A, Kamble V. Comparison of predictive algorithms: backpropagation, SVM, LSTM and Kalman Filter for stock market. In: 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE; 2019. p. 228–234.

[8] Chen K, Zhou Y, Dai F. A LSTM-based method for stock returns prediction: A case study of China stock market. In: 2015 IEEE international conference on big data (big data). IEEE; 2015. p. 2823–2824.

[9] Roondiwala M, Patel H, Varma S. Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*. 2017;6(4):1754–1756.

[10] Yu P, Yan X. Stock price prediction based on deep neural networks. *Neural Computing and Applications*. 2020;32(6):1609–1628. doi: 10.1007/s00521-019-04212-x.

[11] ao P, Zhang R, Yang X. The application of stock index price prediction with neural network. *Mathematical and Computational Applications*. 2020;25(3):53. doi: 10.3390/mca25030053.
[12] Shahi TB, Shrestha A, Neupane A, Guo W. Stock price forecasting with deep learning: A comparative study. *Mathematics*. 2020;8(9):1441. doi: 10.3390/math8091441.
[13] Kara Y, Boyacioglu MA, Baykan ÖK. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert systems with Applications*. 2011;38(5):5311–5319. doi: 10.1016/j.eswa.2010.10.027.
[14] Schöneburg E. Stock price prediction using neural networks: A project report. *Neurocomputing*. 1990;2(1):17–27. doi: 10.1016/0925-2312(90)90013-H.
[15] Kohara K, Ishikawa T, Fukuhara Y, Nakamura Y. Stock price prediction using prior knowledge and neural networks. *Intelligent Systems in Accounting, Finance & Management*. 1997;6(1):11–22. doi: 10.1002/(SICI)1099-1174(199703)6:1<11::AID-ISAF115>3.0.CO;2-3.
[16] Adebiyi AA, Ayo CK, Adebiyi M, Otokiti SO. Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*. 2012;3(1).
[17] Selvin S, Vinayakumar R, Gopalakrishnan E, Menon VK, Soman K. Stock price prediction using LSTM, RNN and CNN-sliding window model. In: 2017 international conference on advances in computing, communications and informatics (icacci). IEEE; 2017. p. 1643–1647.
[18] u B, Wang L, Wang S, Zeng YR. Forecasting the US oil markets based on social media information during the COVID-19 pandemic. *Energy*. 2021; 226:120403. doi: 10.1016/j.energy.2021.120403.
[19] R. Goonatilake and S. Herath, The volatility of the stock market and news, International Research Journal of Finance and Economics, 2007, 11: 53-65.
[20] Spandan Ghose Chowdhury, Soham Routh, Satyajit Chakrabarti, News Analytics and Sentiment Analysis to Predict Stock Price Trends, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3595-3604.
[21] https://news.google.com/
[22] https://finance.yahoo.com/
[23] https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)
[24] https://hex.tech/use-cases/sentiment-analysis/vader-sentiment-analysis/
[25] https://en.wikipedia.org/wiki/Support_vector_machine
[26] https://en.wikipedia.org/wiki/Random_forest
[27] https://en.wikipedia.org/wiki/Data_mining
[28] https://en.wikipedia.org/wiki/Sentiment_analysis