Predicting Stock Market Price Using Sentiment Analysis

Introduction

Conventional time-series analysis methods include autoregressive model (AR), moving average model (MA), autoregressive and moving average model (ARMA) and autoregressive integrate moving average model (ARIMA). All these approaches mainly focus on the time series itself, while ignoring other influencing factors such as the context information. Specifically, they assume the previous data and the later data as independent and dependent variables, respectively, aiming to obtain the quantitative relationship between them. Moreover, these methods often require some assumptions and pre-knowledge, such as the underlying data distribution, valid ranges for various parameters and their connections.

Recently, financial field has been widely using machine learning models in time-series analysis. Support vector regression (SVR) and artificial neural networks (ANNs) both gained considerable results. In addition, deep learning becomes a new trend of machine learning, due to its excellent ability to map nonlinear relationships and adopt limited background knowledge. Deep learning has powerful data processing capabilities that can solve the problems caused by the complexity of financial time series. Therefore, the combination of deep learning and finance has very broad prospects, but the work in this area is not enough.

Literature Review

In the recent review, X. Li, et al proposed a baseline research in which they build up a stock prediction system and propose an approach that first and foremost, converts historical prices into technical indicators that summarize aspects of the price information, and models news sentiments by using different sentiment dictionaries and represents textual news articles by sentiment vectors, secondly, constructs a two-layer LSTM neural network to learn the sequential information within market snapshots series, lastly, constructs a fully connected neural network to make stock trend predictions [1].

I. K. Nti, et al, investigated the potential of public sentiment attitudes (positive vs. negative) and sentiment emotions (joy, sadness and more) extracted from web financial news, tweets, forum discussion and Google trends in predicting stock price movements, using MLP-ANN [2].

A. E. O. Carosia, et al, presented the impacts of the predominant sentiment in the social media Twitter in the Brazilian stock market movements. Among the Machine Learning techniques evaluated, the best technique for SA in Portuguese was the MLPs, which presented better results regarding accuracy and F1-score. This study also showed that it is possible to verify an association between the predominant sentiment in social media and stock market movements within three perspectives: absolute number of tweets sentiments, sentiments weighted by FAVs and sentiments weighted by number of RTs [3].

A. Mittal, et al, have investigated the causative relation between public mood as measured from a large scale collection of tweets from twitter.com and the DJIA values. Our results show that firstly public mood can indeed be captured from the large-scale Twitter feeds by means of simple natural language processing techniques, as indicated by the responses towards a variety of socio-cultural events during the year 2009. Secondly, among the observed dimensions of moods, only calmness and happiness are Granger causative of the DJIA by 3-4 days [4].

Thomas Renault provide some guidelines to help researchers in finance in deriving quantitative sentiment indicators from textual content published on social media [5].

Title	Journal/Published by	Dataset	Methodology	Result
Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong	Information Processing and Management	Hong Kong Stock Exchange data using four different sentiment dictionaries	Author propose an approach that 1) converts historical prices into technical indicators that summarize aspects of the price information, and models news sentiments by using different sentiment dictionaries and represents textual news articles by sentiment vectors, 2) constructs a two-layer LSTM neural network to learn the sequential information within market snapshots series, 3) constructs a fully connected neural network to make stock predictions.	Support Vector Machine (SVM) : F1 score = 0.326 Multiple Kernel Learning (MKL): F1 score = 0.214 LSTM: F1 score = 0.278
Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana	Applied Computer Systems	Historical stock prices of three (3) companies (GCB, MTNGH and TOTAL) listed on GSE, which were downloaded from the official website of GSE. The second was unstructured datasets which included tweets, web news, and post-forum. The	The Multi-Layer Perceptron (MLP) ANN algorithm was adopted for the current study due to its efficiency and effectiveness in predicting the financial market	1-Day Ahead prediction: Accuracy: 70% 7-Day Ahead prediction: Accuracy: 73% 30-Day Ahead prediction: Accuracy: 75%

		tweet data were collected from Twitter, using a Twitter Search API Tweepy. The third dataset was Google trends (DGtrends), a service provided by Google, which enables anyone to find out the volume of search on any topic.		
Stock Prediction Using Twitter Sentiment Analysis	Stanford University	DJIA values from June 2009 to December 2009. Publically available Twitter Dataset.	Cleaved into two phases: 1) Sentiment Analysis: It comprises of multiple stages: • Word list generation • Tweet Filtering • Daily score Computation • Score Mapping 2) Model Learning and Prediction • Prediction Models: SVM, Logistic Regression • Prediction Model: Linear Regression, SOFNN	1) Linear Regression: • MAPE = 7.78% max • Direction = 71.11% max 2) SOFNN: • MAPE = 11.78% max • Direction = 75.56% 3) SVM: • Direction = 59.75% 4) Logistic Regression: • Direction = 60%
Analyzing the Brazilian Financial Market through Portuguese Sentiment	Applied Artificial Intelligence	Brazilian financial market values from October 2018 to December 2018 Twitter tweets from October	• Implementation of a Sentiment Analysis module for the Portuguese language.	Accuracy of Trained model are quoted below: Naïve Bayes = 80.6%

Analysis in Social Media		2018 to December 2018	• Classifying the stock movement by apply different classification models such as Naïve Bayes, SVM, Multilayer Perceptron and, Maximum Entropy	Maximum entropy = 84.1% SVM = 79.7% MLP = 84.8%
Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages	Digital Finance, Springer	150 million StockTwits messages are extracted from StockTwits Application Programming Interface. StockTwit is microblogging platform where users can share ideas and opinions about the stock market. Stock Returns of five biggest U.S. listed stocks: Apple, Amazon, Facebook, Google and Microsoft.	Author has proposed approach to evaluate the performance of a wide range of pre-processing methods and machine learning algorithms for sentiment analysis in finance. Different classification models employed to classify the sentiments are listed below: • Multinomial Naive Bayes • Maximum entropy • Support vector machine • Random forest • Multilayer perceptron	Accuracy of different models are listed below: • Multinomial Naive Bayes = 73.568 • Maximum entropy = 74.451 • Support vector machine = 74.292 • Random forest = 71.665 • Multilayer perceptron = 73.829

Proposed Approach

Proposed approach targets to predict the t^{th} day Stock Price value depending upon $(t-1)^{th}$ day sentiment score of news headlines. Proposed methodology comprises of two segments: first is sentiment analysis, and second is model fitting for prediction of future values and projections.

Sentiment classification techniques can be mainly divided into Machine Learning approach and Lexicon based approach.

Machine learning techniques that are applied in the field of sentiment analysis can be divided as supervised and unsupervised learning methods.

- Unsupervised Learning: Unsupervised learning has no explicit target output associated with input, and it is learning through observation. The goal is to have the machine learn without giving any explicit instruction.
- Supervised Learning: Supervised learning is one that makes use of known dataset to make the prediction of output result. Supervised learning requires two sets of

documents: training set and test set. For learning different properties of documents, training set is used and for evaluating the performance classifier test set is used.

For sentiment analysis lexicon based approach is robust that result in good cross-domain performance. This method is based on the assumption that the sum of the sentiment orientation of each word makes contextual sentiment orientation.

- Dictionary Based approach: This approach use predefined dictionary of words where each word is associated with a specific sentiment polarity strength. Feeling of people such as happy, sad or depressed can be found out by comparing word against lexicons from dictionaries.
- Corpus based approach: Corpus based approach try to find co-occurrence patterns of words to determine their sentiments. This approach is based on seeding list of opinion words and then find another opinion words which have similar context. This method is used to assign happiness factor of words depending on frequency of their occurrences in “happy” or “sad” blog post.

Once sentiment scores are obtained, obtained score are correlated with the S&P 500 index of stock prices. Depending upon correlation score, further approaches were prepared.

Once data frame is collected and cleaned, it becomes ready for analysis. For analysis perspective, first and foremost thing done to the dataset is data split. Basically, training and testing datasets are prepared in data split by differentiating dataset into two distinctive parts depending upon the differentiating ratio.

- Training Dataset: The dataset used for preparing the model (loads and predispositions on account of Neural Network) is considered as Training Dataset. The model sees and gains from this data.
- Testing Dataset: The test dataset is basically employed for maintaining quality level which further can be used to evaluate the model. It is simply used once a model is completely prepared. The test dataset is once utilized in whole analysis to assess contending models.

Once the data split is performed successfully, data frame are ready for the model fitting. Different regression models are prepared and there accuracy and error rate is calculated. Also, few Deep learning models are also employed to prepare the required results and control the error rate. All these prepared model are analysed and there error matrix is prepared for better understanding. Mean Absolute Error and Root Mean Square Error is employed for deducing the error rate of the models.

Results

Proposed approach is cleaved into two parts, sentiment analysis and model training.

As mentioned in proposed approach section of this manuscript, sentiment analysis is done using dictionary based approach. VADER dictionary which is prepared by MIT, is employed to for deducing the sentiment scores.

Two types of datasets are employed for sentiment analysis:

- ReditNews Dataset: It incorporates top trending news of the day from the period of 2008 to 2016. Dataset contains attribute date and Headlines.
- Top 25 News Dataset: It incorporates the top 25 headlines of the day for the time scale of 2008 to 2016. Dataset contains 26 attributes, 25 for each headline and 1 for Date.

- S&P 500 Index: It contains Date, Open, Close, High, Low, and Adj. Close fields, which contains stock price values for each day.

As part of data pre-processing, all the special characters were removed from the headline text, and all the news headlines of same date are combined as a single data point. Then, corresponding to headline date, S&P 500 index stock prices are fetched from the S&P 500 index dataset retrieved from yahoo finance. Moreover, all the empty data points in Stock price field are filled with the mean of available stock prices. All these steps has prepared a sorted data frame upon which sentiment analysis can be performed, and required model can be trained.

For performing sentiment analysis, predefined libraries of NLP in python are imported and specifically VADER library of NLP is incorporated for further analysis. Using these libraries, sentiment scores are prepared. Sentiment scores are obtained in four different fields, i.e. Positive, Negative, Neutral, and Compound. Positive, negative, and neutral field will record positive, negative and neutral sentiment score, and Compound field is the sum of previous three and further normalized to in the range of [-1, 1].

After preparing sentiment scores, next step was of model training for accurate and precise prediction. But before actual model training, correlation was calculated among the different sentiment field with stock prices. Below cited table gives the correlation score among different fields.

Table 1 Correlation Matrix

	Compound	**Neutral**	**Negative**	**Positive**
Top Headline News Dataset	0.012436	0.244070	-0.228484	0.061837
RedditNews Dataset	0.00980	0.15620	-0.152699	-0.023876

After observing correlation results, it was clear that regression model will not fit properly to the dataset employed. But due to correlation results of Neutral and Negative which is actually not that bad, regression model is can prepared. So, both prepared data frames are as passed for regression model fitting. Models employed are Linear Regression, Support Vector Regression, Decision Tree Regression, and Gradient Boasting. Results obtained after predicting the stock values are quoted in Table 2 and Table 3.

Since, Mean Absolute Error is quite high for all the regression model, there is need of deep learning model to control the error rate and predicting accurate and precise stock prices. So, Artificial Neural Network (ANN) and Long Short Term Model (LSTM) is trained with above processed data frame. Correlation matrix when observed parallel to results of both types of models, i.e. regression models, and deep learning models, gives the reason why regression models performed worst and deep learning models are best fit models. Also, dataset employed also differs the results. Reddit News dataset give best results only on ANN model, whereas News Headline dataset gives best results for Extreme Gradient Boosting Algorithm. Results obtained from both datasets are quoted in Table 2 and 3.

Table 2 Results for RedditNews Dataset

	Mean Absolute Error	**Root Mean Square Error**
Test for Linear Regression	388.736436	474.710369
Test for Decision Tree Regressor	390.907979	475.075729

Test for Extreme Gradient Boosting	389.042819	472.349167
Test for LSTM	214.683640	318.972006
Test for ANN	0.161128	0.225732

Table 3 Results for News Headline Dataset

	Mean Absolute Error	**Root Mean Square Error**
Test for Linear Regression	600.018919	600.018939
Test for Decision Tree Regressor	0.033534	0.130441
Test for Extreme Gradient Boosting	0.021036	0.098504
Test for LSTM	17.905527	38.478966
Test for ANN	0.226042	0.272522

Conclusion

Unlike the conventional stock market prediction systems, novel approach combines the sentiments of common people through the news feeds and stock price data to predict the behaviour of stock market. The Reddit news feed of the top headlines are obtained and sentiment polarity of the news sentences are calculated for the prediction of news, whether it is positive, negative or neutral. Depending upon the sentiment score of particular day, next day's stock price is mapped and model is trained according to that approach. After analysing the results obtained, it is clearly observable, deep learning model i.e. ANN, is best fit for the forecasting for Reddit News dataset, as the MAE of ANN is 0.161. Also, for News Headline dataset Extreme Gradient Boosting and ANN model gives best results, as the MAE for Extreme Gradient Boosting, Decision Tree Regressor and ANN are 0.021036, 0.033534, and 0.226042 respectively. However, proposed approach can be improved by incorporating sentiment score of different dictionaries for sentiment analysis, and employing other ensemble models for making the prediction.

References

[1] Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, 102212.

[2] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. *Applied Computer Systems*, *25*(1), 33-42.

[3] Carosia, A. E. O., Coelho, G. P., & Silva, A. E. A. (2020). Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, *34*(1), 1-19.

[4] Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, *15*.

[5] Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance*, *2*(1), 1-13.