Forecasting Stock Market Alternations Using Social Media Sentiment Analysis and Regression Techniques

Christina Saravanos[1] and Andreas Kanavos[2(✉)]

[1] Computer Engineering and Informatics Department, University of Patras, Patras, Greece
saravanou@ceid.upatras.gr
[2] Department of Informatics, Ionian University, Corfu, Greece
akanavos@ionio.gr

Abstract. In recent years, the public opinion is swayed by online social, media and news platforms, such as Twitter, podcasts, and streaming news broadcasts. The public opinion can alter the outcome of various social-economic events, e.g., the volatility of the stock market. This paper presents an overview of forecasting the volatility of the indices of several companies in the U.S. stock market while considering the sentiment and features extracted from the metadata of a tweet and its author's social activity and network. The daily changes in the prices of an index in the U.S. stock market were estimated by applying several regression techniques. The results indicate a strong correlation between the approximated closing prices of the stocks in the U.S. stock market, the sentiment along with the features extracted from a tweet, and its author's activity and network. Finally, the obtained results indicate that the number of attributes did not impact the performance of the applied regression techniques.

Keywords: Knowledge Extraction · Stock Forecasting · Sentiment Analysis · Social Media Analysis · Natural Language Processing (NLP) · Twitter · Regression Techniques

1 Introduction

Nowadays, public opinion is swayed by several online social and media platforms, e.g., social networks and streaming services. Most social networks, such as Facebook, allow their users to upload and comment on a post regarding an event. On the other hand, Twitter endorses its users to upload short-length messages or posts. These posts are referred to as tweets. A tweet comprises 140 characters. Twitter, additionally, endorses its users to interact by replying to or *retweeting* a tweet. Therefore, a tweet will diffuse from its author's social network to the respective networks of its author.

Published by Springer Nature Switzerland AG 2023
I. Maglogiannis et al. (Eds.): AIAI 2023 Workshops, IFIP AICT 677, pp. 335–346, 2023.
https://doi.org/10.1007/978-3-031-34171-7_27

Furthermore, tweets cover a wide range of subjects. A tweet can either refer to newsworthy events or reflect the opinion or the emotion of their authors regarding a product or service, e.g., the latest MacBook. Most tweets indicate the feelings of their authors towards a subject via certain words, hashtags, punctuation marks, or emojis. As a tweet spreads through millions of social networks, the emotion or the opinion of its author will impact the feelings or the emotion of several millions of users on Twitter.

In recent years, however, several novel techniques have emerged that seek to estimate the fluctuations in the daily prices of an index by relying on the correlation between its historical prices and the sentiment extracted from the respective tweets. The volatility of the index is estimated by either using regression techniques, e.g., linear regression, or deep neural networks such as RNNs [5,8,9,12,17,24].

This paper aims to approximate the volatility of the indices of several companies in the U.S. stock market by relying on the correlation between the daily prices of the indexes, the sentiment, and several features extracted from the respective tweets. The features were extracted from the metadata of a tweet and its author's presence on Twitter. The elicited features indicate the impact of a tweet on its author's social network. In other words, the extracted features hint at the initial impact of the tweet's sentiments on its author's friends and followers. As the tweet diffuses into the social networks of hundreds of users it will impact the emotion and opinions of thousands. The volatility of an index was estimated via regression techniques. The results initially suggest that the daily closing prices and the sentiment are strongly correlated. In addition, the results indicate that the estimated closing prices depend on the previously-mentioned features. Finally, the results suggest that the volume of tweets plays a rather significant role in predicting future closing prices.

The remainder of this paper is organized as follows: Sect. 2 briefly depicts the novel techniques that have recently emerged to forecast the volatility of the stock market by relying on the correlation between the daily closing prices of an index and the emotion extracted from the respective tweets. Section 3 elaborates on the presented scheme, particularly on the extracted features and employed regression techniques. Section 4 comments on the experimental setup and the results that emerged while conducting the experiments. Finally, Sect. 5 draws the final conclusions.

2 Related Work

The volatility of an index in the stock market is affected by several distinct factors, such as the public opinion, the news, and trends. Therefore, predicting the volatility of the stock market is considered an underlying problem.

As the prominence of social network platforms continues to grow, innovative paradigms have been introduced which seek to estimate the volatility of the stock market by relying on regression techniques and the correlation between past prices of an index and the sentiment extracted from several online posts

which mention the respective company. These posts have been posted on a social media platform. In [26,30], the volatility of the Indian and U.S. stock markets was, correspondingly, forecast by, initially, computing the correlation between the past prices of the NSE and the DJIAI indices along the public sentiment extracted from the respective tweets.

Moreover, [1,10,27,31] aspired to predict the future movements of the FTS100, BSE, and the DIJAI indices, respectively, by relying on the interrelationship of the former alternations of the indices and the emotion elicited from the respective tweets as well as the Granger Causality Test. In addition, the future volatility of the stock market in [27,31] was estimated by employing auto-regression methods such as the Auto-Regressive Moving Average (ARIMA). In [13], however, Gupta et al. proposed a novel paradigm that sought to forecast the volatility of the U.S. stock market by relying on the correlation between the current daily prices of an index and the sentiment excerpted from the respective tweets.

Furthermore, the scheme introduced in [15] extracted features from several social media platforms such as Twitter and Google, the sentiment of thousands of tweets to predict the volatility of the U.S. stock market via a novel regression technique, the Delta Naive Bayes. On the other hand, [18] aspired to estimate possible changes in the future prices of several indices of the U.S. stock market by employing the correlation between the respective previous prices and the emotion extracted from the respective tweets along with several conventional regression techniques, such as RF, KNN, and SVR.

3 Overview of the Presented Scheme

This section elaborates on the forecasting of the volatility in the stock market. The presented scheme comprises three distinct modules. The first module extracts the emotion and several features derived from the tweet and its author's metadata. The second one elicits attributes from public information regarding a company's stock. The tweets and the stock market information were extracted by employing the Twitter and Yahoo! Finance API. Finally, the third module estimates the daily closing price of a company depending on the features provided by the first two components and by employing a regression technique.

3.1 Feature Extraction

This subsection elaborates on the sentiment and the features obtained from the first two modules of the presented scheme. The depicted paradigm relies on the tweets and the daily stock market information acquired from Twitter and Yahoo! Finance API, a platform focused on providing financial news.

3.1.1 Extracting the Sentiment and Features from A Tweet

The first module seeks to extract several features from the content of a tweet and its author's overall presence on Twitter. The attributes obtained from the tweet's metadata and its authors' social activity, as presented in Table 1.

Table 1. Features extracted from Tweet's Metadata and Tweet's Author Profile.

Tweet's Metadata			
IRT	Is Retweet	IRP	Is Reply To
NH	# Hashtags	NU	# URLs
NMU	# Mentioned Users		
Tweet's Author Profile			
NTw	# Tweets	NFoll	# Followers
NFr	# Friends	NL	# Lists
NTU	# URLs used	NTMU	# Users mentioned
NTH	# Hashtags used	NFT	# Favorite Tweets
NTC	# Conversations participated		

Moreover, the first module aims to extract the sentiment from each particular tweet. *Sentiment analysis* consists of a group of techniques that aim to excerpt the sentiment of a tweet towards a product, event, or organization. Sentiment analysis seeks to classify a tweet into one of three sentiments: positive, negative, or neutral by applying either lexicon-based approaches, such as the Natural Language Processing Kit's (NLTK) Vader Analyzer, or supervised learning models such as in [3,16].

NLTK's Vader Analyzer is a lexicon- and rule-based sentiment analysis tool that extracts the sentiment of the tweets via a combination of sentiment lexicons. The sentiment lexicons are a list of lexical features labeled as either positive, negative, or neutral based on their semantic orientation. A polarity score is computed based on the sentiment lexicons [6,23,29,33].

3.1.2 Extracting Stock Market Features

Several stock market features were obtained from the daily public information obtained from Yahoo! Finance and correspond to the number of shares, the opening, closing, and high and low prices of an index in the U.S. stock market. Two additional features, the high/low percentage, and the percentage change, were computed based on the previously-mentioned attributes. The former corresponds to the daily changes between the high and low prices, whereas the latter coincides with the daily alternations between the opening and closing prices of an index.

The two additional features are defined as follows:

$$\text{High/Low Percentage} = \frac{High - Low}{Low} * 100 \tag{1}$$

$$\text{Percentage Change} = \frac{Close - Open}{Open} * 100 \tag{2}$$

where *High*, *Low*, *Close* and *Open* denote the high, low, closing and opening prices of an index in the stock market, respectively.

3.2 Regression Techniques

The volatility of the indices was approximated by employing several regression techniques. Regression models seek to predict the value of the dependent variable, such as the daily closing price, by relying on the independent variables, e.g., the features extracted from the dataset.

Linear Regression (LR) aims to find a relationship between the independent and dependent variables by fitting the data into a linear equation. The relationship between the independent and dependent variables hints at a strong association between the two. Two types of linear regression, *single* and *multivariate linear regression*, are employed depending on the size of the independent variables and the complexity of their relationship with the dependent variables. Both types compute the values of the dependent variable to fit the respective independent or predictor values into a linear model [11,22,25].

Moreover, the Support Vector Regression (SVR) algorithm is a non-linear regression approach that seeks to forecast future data points by employing historical samples provided by the dataset. In a high-dimensional feature space, a linear function, referred to as the SVR function aims to create the non-linear relationship between the input and output data, or the predictor and predicted variables [7,14,20].

On the other hand, Decision Trees (DTs) are non-parametric models applied to solve either classification or regression problems. A DT consists of a root or parent node, a set of internal and terminal nodes. The dataset is initially allocated at the root node and is subsequently segmented in every internal node. The internal nodes correspond to predictors. The classes or target values that have emerged from several subsequent internal nodes, are assigned to the terminal nodes. The DT regressor relies on applying binary recursive partitioning, i.e., a technique that iteratively splits the dataset into several groups. The regressor partitions the root node into several binary pieces. A binary piece is assigned to a child node. The child node corresponds to an internal node. The regression then selects the child node that either minimizes the sum of squared deviations which emerged from splitting the dataset into two parts or corresponds to the feature with the lowest impurity index. The impurity of a DT is calculated via the *Gini Index*. The dataset splitting is applied on several internal nodes and ends when a node reaches a minimum node size; this node corresponds to a terminal node [19,32].

Random Forest (RF) is a regression technique that relies on the result and the performance of several DT regressors. The RF scheme is initialized by constructing several hundreds or thousands of de-correlated DT regressors. The DTs are implemented via a randomized subset of predictors to create the random forest. The output of RF is equal to the mean value of the outputs of the DTs. Therefore, RF provides more accurate results in forecasting the future values of a target node than DT. The RF depends heavily on an approach often referred to as *bagging* or *bootstrap aggregation*. *Bagging* seeks to reduce the variance often associated with a regression model to increase the performance of the RF. In addition, bagging decreases the correlation between the implemented DTs by

randomly selecting a feature subset from each DT. Therefore, a feature may be employed several times, while others are never selected to train the RF. The random selection of feature subsets is attributed to as *bootstrap*. The selected bootstrap samples are fed into the RF scheme and will result in creating a smaller subset of regressors. The aggregated output of the RF emerges by calculating the mean output of the selected subset of DTs [19,28].

Finally, the AdaBoost model is an ensemble, iterative paradigm that seeks to create a robust learning-based approach by linearly combining several weak learning algorithms. The learning methods are alluded to as *base learners*. In each iteration, a base learner is called and constructs a weak learning-based paradigm. Thereupon, a weight coefficient is assigned to every learning-based model. The accuracy of the base learners increases by re-calculating their respective weight coefficients. The output of the AdaBoost model emerges by summing the outputs of the learners multiplied by their corresponding weight coefficients [2,4,21].

4 Experimental Setup and Results

This section elaborates on the experimental setup and the results that emerged from conducting the experiments. The presented scheme aims to estimate the daily prices of the indices mentioned in Table 2 by relying on the correlation between the features and via the regression techniques described in Sect. 3.

Table 2. The names, the indices and the three most frequent hashtags obtained from the tweets posted from March 22^{nd} to March 31^{st}, 2021 of the companies whose volatility is predicted by the presented scheme.

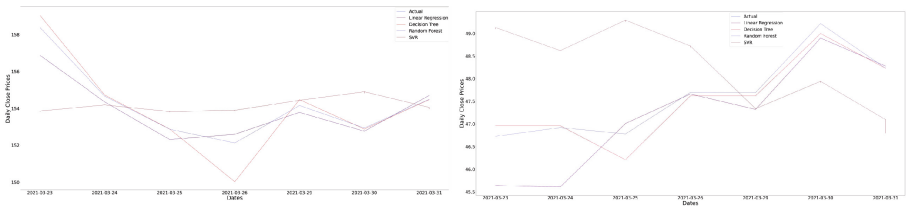
Company	Stock Index	Three Most Frequent Hashtags
Amazon	AMZN	#amazon, #deals, #sales
Apple	APPL	#Apple, #AppleMusic, #iphone
Delta	DAL	#DAL1669, #AIRBUS, #DAL1038
Google	GOOGL	#google, #DoodleForGoogle, #chrome
Microsoft	MSFT	#windows, #microsoft, #Azure

4.1 Datasets

The daily stock market information of the indices and several thousands of tweets that mention the companies depicted in Table 2 were respectively extracted from the Yahoo! Finance and the Twitter API from March 22^{nd} to March 31^{st}, 2021. The U.S. stock market is closed on the weekends. In other words, the tweets posted on the respective dates were discarded. The presented scheme aims to forecast the volatility of the indices of several companies; therefore, several datasets were created.

4.2 Estimating the Changes of the Daily Closing Prices in the Stock Market

The daily closing prices of the indices of the companies presented in Table 2 from March 22^{nd} to March 31^{st}, 2021 were predicted by applying the techniques mentioned in Sect. 3. Figure 1 illustrates the actual and the respective closing prices of the indices that correspond to Amazon and Delta that emerged from employing the previously - mentioned regression techniques.



(a) Amazon (b) Delta

Fig. 1. The actual and estimated daily closing prices of the indices of (a) Amazon and (b) Delta in the U.S. stock markey from March 22^{nd} to March 31^{st}, 2021.

The results depicted in Fig. 1 suggest similar movements between the actual and the estimated daily closing prices of the indices. In several cases, the results indicate that the actual daily closing prices of the indices coincided with the respective estimated prices. Nonetheless, the closing prices that emerged via the SVR deviated from the actual prices of the companies' indices.

Table 3 presents several metrics referring to the total number of tweets and their overall diffusion on Twitter from March 22^{nd} to March 31^{st}, 2021. Figure 1 and Table 3 suggest that as the number of total tweets and their respective diffusion on Twitter increases, the predicted prices coincide with the actual closing prices of the index on Twitter. In a similar manner, a low volume and dispersion of tweets led to the estimated closing prices diverging from the corresponding actual prices in the stock market. Therefore, the volume and diffusion of the tweets play a rather significant role in forecasting the volatility of an index.

The estimated daily closing prices of an index deviated from the respective actual prices due to the high number of features. In several cases, the value of an attribute was equal to zero. Consequently, the sparse dataset led to the curse of dimensionality which led the models to overfit.

4.3 RMSE and MAPE

The accuracy of a regression technique relies on calculating the error between the actual and estimated value of an observation. In most real-world applications,

Table 3. The total number of tweets, retweeted tweets, replied tweets, followers and friends of the tweets' authors extracted from Twitter between March, 22^{nd} to March, 31^{st}, 2021.

Company	Tweets	Retweeted Tweets	Replied Tweets	Followers of Tweets' Authors	Friends of Tweets' Authors
Amazon	51,053	10,817	1,553	443,059,418	577,441
Apple	31,421	23,233	641	181,102,724	636,461
Delta	3,277	1,196	196	77,852,621	66,557
Google	113,659	87,374	1,305	553,284,947	3,085,156
Microsoft	10,065	5,682	283	121,683,430	213,395

the error relies on either the Root Mean Square Error (RMSE) or the Mean Absolute Percentage Error (MAPE), which are defined as:

$$RMSE = \frac{1}{\sum_{i=0}^{N} x_i} * \sqrt{\frac{\sum_{i=0}^{N}(x_i - \hat{x_i}}{N}} \tag{3}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{x_i - \hat{x_i}}{x_i} \right| \tag{4}$$

where N, x_i, $\hat{x_i}$ denote the total number of observations, the actual and the estimated value, that emerged from the techniques depicted in Sect. 4 of the i^{th} observation, correspondingly. The RMSE indicates the deviation of the estimated price from the respective actual daily closing price of an index, whereas the MAPE, an accuracy metric, hints at the percentage of change between the predicted and actual closing prices of an index in the market.

Tables 4 and 5 present the RMSE and MAPE which emerged from estimating the indices depicted in Table 2 from March 22^{nd} to March 31^{st}, 2021. The depicted results suggest a similar divergence to the results illustrated in Fig. 1. These results indicate that the MAPE is a better accuracy metric than the RMSE since it points to the actual fluctuation between the actual and predicted

Table 4. RMSE between the companies' actual and approximated close prices in the stock market from March 22^{nd} to March 31^{st}, 2021.

Company	LR	DT	RF	AdaBoost	SVR
Amazon	1.1237e−09	0.4285	0.6720	0.3418	1.5415
Apple	3.2012e−09	0.3149	0.4568	0.2736	2.7477
Delta	1.1254e−10	0.9379	0.7471	1.0724	2.2164
Google	6.2843e−10	0.5052	0.1628	0.2266	1.0376
Microsoft	3.4393e−10	0.5084	0.4738	0.7083	3.2809

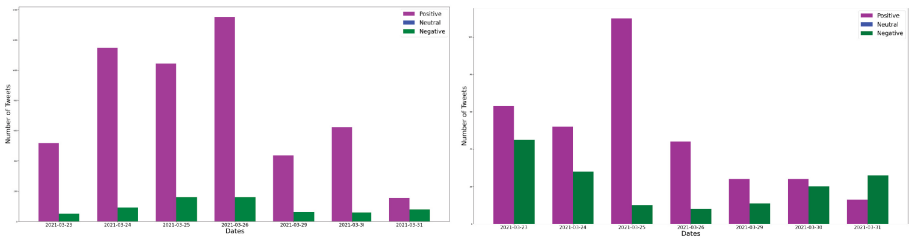
Table 5. MAPE between the companies' actual and approximated close prices in the stock market from March 22^{nd} to March 31^{st}, 2021.

Company	LR	DT	RF	AdaBoost	SVR
Amazon	5.6796e−10	0.2330	0.3637	0.1891	0.8425
Apple	2.1637e−09	0.2201	0.3470	0.2042	2.0782
Delta	2.0135e−10	1.2442	1.1789	1.7419	4.1396
Google	5.7707e−10	0.4388	0.1421	0.2127	0.8821
Microsoft	1.2400e−10	0.1619	0.1364	0.2018	1.1091

closing prices of the indices. Finally, the numerical values of the MAPE point to a similar deviation of the actual and approximated closing prices illustrated in Fig. 1.

4.4 Correlation Between Daily Closing Prices and Sentiment Extracted from Tweets

While conducting the experiments, polarity scores, which indicate the sentiment, were extracted from the tweets. The tweets were classified as either positive, neutral, or negative based on the value of their respective polarity scores. Figure 2 illustrates the daily number of positive, neutral, and negative tweets that were posted on Twitter from March 22^{nd} to March 31^{st}, 2021.



(a) Amazon (b) Delta

Fig. 2. The daily number of positive, neutral and negative tweets that mention (a) Amazon and (b) Delta posted on Twitter from March 22^{nd} to March 31^{st}, 2021.

Figure 2 and Table 3 cite the overall sentiment and diffusion of the tweets in the social network. The depicted information indicates the impact of the tweets on their authors' friends and followers on Twitter. If a tweet spreads throughout Twitter, then its sentiment will influence the emotion or opinions of thousands. Therefore, the public sentiment will be swayed positively or negatively.

The results and the information depicted in Figs. 1 and 2 respectively suggest that the sentiment extracted from the tweets and the closing prices of the stock

market indices are strongly correlated. The illuminated results hint that as the number of positive tweets that mention a company, increases, its stock will be overbought. Therefore, the closing price of the company's index will fall in the stock market. Similarly, the increase in negative tweets that mention a company, will lead to the company's stock being oversold. Thus, the closing price of the company's index will rise.

5 Conclusions and Future Work

This paper presented a scheme that forecasts the volatility of the daily closing prices of stocks, of several companies, in the U.S. stock market via several regression techniques. The advocated paradigm, additionally, relies on the correlation between the past daily closing prices of the respective companies' stocks, the emotion, and several features extracted from the metadata and text of a tweet as well as from the social activity and network of its author on Twitter.

The advocated scheme seeks to forecast the volatility of the stock market by combining the respective past prices, the emotion, and several features regarding the tweet and its author. The changes in the daily closing prices of a company's stock in the U.S. market are approximated by applying several regression techniques. The accuracy of the estimated and the actual closing prices was calculated via the RMSE and the MAPE. The RMSE and the MAPE between the actual and the estimated closing prices that emerged from applying linear regression and SVR are rather low. In other words, the difference between the actual and the estimated prices are close to zero. On the other hand, the RMSE and the MAPE obtained from applying DT, RF, and AdaBoost regression are quite large indicating the presence of noise and curse of dimensionality which occurred due to the vast number of features.

Moreover, the obtained results hint at a strong correlation between the volatility of a company's stock in the market and the sentiment extracted from the tweets in which the companies are mentioned. The acquired results, additionally, point out that the estimated daily changes in the closing prices of a company's stock are correlated to the features extracted from the tweet and its author's social activity on the social media platform.

Regarding the directions for future research, the volatility of the daily closing prices, of a company's stock, can be approximated via the correlation between the respective past prices, the sentiment, and the subjectivity of a tweet. Furthermore, the daily changes in the closing prices can be estimated by applying DL frameworks, such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs). In addition, the changeability of the prices in the stock market can be estimated by taking into account the current trends of Twitter, or another social media platform or search engine, such as Google. Finally, features obtained from the social profiles of a user from several other social networks, such as LinkedIn, Facebook, and Reddit, can be concatenated to forecast future changes in a company's price.

Acknowledgement. This research was co-financed by the European Union and Greek national funds through the "Competitiveness, Entrepreneurship and Innovation" Operational Programme 2014–2020, under the Call "Support for regional excellence"; project title: "Intelligent Research Infrastructure for Shipping, Transport and Supply Chain - ENIRISST+"; MIS code: 5047041.

References

1. Ahuja, R., Rastogi, H., Choudhuri, A., Garg, B.: Stock market forecast using sentiment analysis. In: 2nd IEEE International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1008–1010 (2015)
2. Ampomah, E.K., Qin, Z., Nyame, G., Botchey, F.E.: Stock market decision support modeling with tree-based adaboost ensemble machine learning models. Informatica (Slovenia) **44**(4) (2020)
3. Baltas, A., Kanavos, A., Tsakalidis, A.K.: An apache spark implementation for sentiment analysis on twitter data. In: 2nd International Workshop on Algorithmic Aspects of Cloud Computing (ALGOCLOUD), vol. 10230, pp. 15–25 (2016)
4. Barrow, D.K., Crone, S.F.: A comparison of adaboost algorithms for time series forecast combination. Int. J. Forecast. **32**(4), 1103–1119 (2016)
5. Bing, L., Chan, K.C.C., Ou, C.X.: Public sentiment analysis in twitter data for prediction of a company's stock price movements. In: 11th IEEE International Conference on e-Business Engineering (ICEBE), pp. 232–239 (2014)
6. Bonta, V., Kumaresh, N., Janardhan, N.: A comprehensive study on lexicon based approaches for sentiment analysis. Asian J. Comput. Sci. Technol. **8**(S2), 1–6 (2019)
7. Chahboun, S., Maaroufi, M.: Performance comparison of support vector regression, random forest and multiple linear regression to forecast the power of photovoltaic panels. In: 9th IEEE International Renewable and Sustainable Energy Conference (IRSEC), pp. 1–4 (2021)
8. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. PeerJ Comput. Sci. **7**, e623 (2021)
9. Das, S., Behera, R.K., Kumar, M., Rath, S.K.: Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia Comput. Sci. **132**, 956–964 (2018)
10. Deveikyte, J., Geman, H., Piccari, C., Provetti, A.: A sentiment analysis approach to the prediction of market volatility. CoRR abs/2012.05906 (2020)
11. Fumo, N., Biswas, M.A.R.: Regression analysis for prediction of residential energy consumption. Renew. Sustain. Energy Rev. **47**, 332–343 (2015)
12. Guo, X., Li, J.: A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. In: 6th IEEE International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 472–477 (2019)
13. Gupta, R., Chen, M.: Sentiment analysis for stock price prediction. In: 3rd IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 213–218 (2020)
14. Hu, J., Gao, P., Yao, Y., Xie, X.: Traffic flow forecasting with particle swarm optimization and support vector regression. In: 17th IEEE International Conference on Intelligent Transportation Systems (ITSC), pp. 2267–2268 (2014)

15. Jin, F., Wang, W., Chakraborty, P., Self, N., Chen, F., Ramakrishnan, N.: Tracking multiple social media for stock market event prediction. In: 17th Industrial Conference on Advances in Data Mining (ICDM), vol. 10357, pp. 16–30 (2017)
16. Kanavos, A., Perikos, I., Hatzilygeroudis, I., Tsakalidis, A.K.: Emotional community detection in social networks. Comput. Electr. Eng. **65**, 449–460 (2018)
17. Kanavos, A., Vonitsanos, G., Mohasseb, A., Mylonas, P.: An entropy-based evaluation for sentiment analysis of stock market prices using twitter data. In: 15th IEEE International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 1–7 (2020)
18. Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H., Alfakeeh, A.S.: Stock market prediction using machine learning classifiers and social media, news. J. Ambient. Intell. Humaniz. Comput. **13**(7), 3433–3456 (2022)
19. Li, Y., et al.: Random forest regression for online capacity estimation of lithium-ion batteries. Appl. Energy **232**, 197–210 (2018)
20. Lin, K., Lin, Q., Zhou, C., Yao, J.: Time series prediction based on linear regression and SVR. In: 3rd IEEE International Conference on Natural Computation (ICNC), pp. 688–691 (2007)
21. Liu, Q., Wang, X., Huang, X., Yin, X.: Prediction model of rock mass class using classification and regression tree integrated adaboost algorithm based on tbm driving data. Tunn. Undergr. Space Technol. **106**, 103595 (2020)
22. Maulud, D.H., Abdulazeez, A.M.: A review on linear regression comprehensive in machine learning. J. Appl. Sci. Technol. Trends **1**(4), 140–147 (2020)
23. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
24. Mittal, A., Goel, A.: Stock prediction using twitter sentiment analysis. Standford University, CS229 15, 2352 (2012)
25. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis. Wiley, Hoboken (2021)
26. Oliveira, N., Cortez, P., Areal, N.: Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In: 3rd ACM International Conference on Web Intelligence, Mining and Semantics (WIMS), p. 31 (2013)
27. Rao, T., Srivastava, S.: Analyzing stock market movements using twitter sentiment analysis. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2012)
28. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M.: Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. **71**, 804–818 (2015)
29. Sahayak, V., Shete, V., Pathan, A.: Sentiment analysis on twitter data. Int. J. Innovative Res. Adv. Eng. (IJIRAE) **2**(1), 178–183 (2015)
30. Sharma, V., Khemnar, R., Kumari, R., Mohan, B.R.: Time series with sentiment analysis for stock price prediction. In: 2nd IEEE International Conference on Intelligent Communication and Computational Techniques (ICCT), pp. 178–181 (2019)
31. Souza, T.T.P., Kolchyna, O., Treleaven, P.C., Aste, T.: Twitter sentiment analysis applied to finance: a case study in the retail industry. CoRR abs/1507.00784 (2015)
32. Xu, M., Watanachaturaporn, P., Varshney, P.K., Arora, M.K.: Decision tree regression for soft classification of remote sensing data. Remote Sens. Environ. **97**(3), 322–336 (2005)
33. Yao, J.: Automated sentiment analysis of text data with nltk. J. Phys. Conf. Ser. **1187**, 052020 (2019)