Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Are Twitter sentiments during COVID-19 pandemic a critical determinant to predict stock market movements? A machine learning approach

Pradyot Ranjan Jena*, Ritanjali Majhi

School of Humanities, Social Sciences and Management, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India

**A R T I C L E   I N F O**

Article history:
Received 18 August 2022
Revised 22 November 2022
Accepted 28 November 2022

Editor by Pradyot Ranjan Jena

Key words:
Sentiment analysis
COVID-19
Stock market
LSTM
Twitter

**A B S T R A C T**

The problem of stock market prediction is a challenging task owing to its complex nature and the numerous indirect factors at play. The sentiments regarding socio-political issues such as wars and pandemics can affect stock prices. The spread of the COVID-19 pandemic continues to take a toll on the economy and fluctuations in sentiment of the concerns about the health impacts of the disease can be captured from the microblogging platform, Twitter. We examined how these sentiments during the Covid-19 pandemic and the health impacts arising from the disease along with other macroeconomic indicators provide useful information to predict the stock indices in a more accurate manner. We developed a machine learning model namely, long-short term memory (LSTM) networks to predict the impact of the Covid-19 induced sentiments on the stock values of different sectors in the United States and India. We did the same predictions using the timeseries statistical models such as autoregressive moving average model and the linear regression model. We then compared the performance of the LSTM and the timeseries statistical models to find that the machine learning model has produced more accurate predictions of the stock indices. The performance of the models across the sectors and between the United States and India are compared to draw economic inferences.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of African Institute of Mathematical Sciences / Next Einstein Initiative. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

**Introduction**

The stock market has been observed to be a volatile signal and predicting these stock market movements has been the focus of analysts. Various models have been developed with the aim of capturing the patterns in stock market prices/indices Chen et al. [12] [10,11,13,34–36]). The most common approach in stock market analysis literature is the Technical analysis [1,2,11] which uses the stock prices or the derivatives of it as input. The underlying assumption is that the stock prices represent the macroeconomic indicators and news. Hence, the information embedded in stock prices is enough to predict the stock market. Macroeconomic variables like Gross Domestic Product, interest rates, currency exchange rates, customer price index, among others have been commonly used for stock indices forecasting [5,43]. Furthermore, text mining techniques are in use to include information such as financial news. However, with the advent of social network analysis more

* Corresponding author.
E-mail address: pradyotjena@nitk.edu.in (P.R. Jena).

https://doi.org/10.1016/j.sciaf.2022.e01480
2468-2276© 2022 The Author(s). Published by Elsevier B.V. on behalf of African Institute of Mathematical Sciences / Next Einstein Initiative. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
recently, using sentiment indices and other derived series as inputs, has proved useful for the stock forecasting ([44]; Barber and Terrance, [3]).

Global Health Emergency of the COVID-19 outbreak was declared by the World Health Organization (WHO) on 30th January 2020 [17,39]. With over 7 million confirmed cases and 400,000 deaths worldwide, the pandemic continues to have impact on the economy [45]. Some of these impacts include the spike in unemployment in the US which is over 20 million, where the prospect of a large number of marginalized people are pushed to poverty. In addition, COVID-19 has significant implication for the performances of the financial markets.

Twitter is an online microblogging platform and acts as a platform where an overall picture of the mood and general sentiment about specific topics can be observed. Twitter users across the world have been closely following the trends of the spread of the pandemic and share their views on the platform. Owing to the lockdown, more people are at home which explains the increased usage of the platform. In this paper, we study to what extent the twitter sentiment during the pandemic can help predict the stock indices in the US and India across a variety of sectors. Furthermore, the trends in healthcare conditions can help for useful information about the performance of the economy. The interpretation of the healthcare trends by people that produce expectations about the performance of the economy is also a crucial feature that we try to leverage.

The previous studies have shown the impact of monetary policy on stock movements. Sirucek [38] has used the DJIA index to represent stock market movement and M2 and M2M aggregates as the money supply to study the co-movement in the US. Li [23] showed how the money supply in Europe during the 2009 crisis had influenced the stock market. They used the Granger causality test and Vector Error Correction Model to shed light on both the short term as well as the long-term inter-relation. Studies have also examined the effect of floating foreign exchange rates on the stock market prices (Mechri et al., [24]). They used the monthly data on stock prices, gold prices, exchange rates and inflation rates for Tunisia from 2002 to 2017. Using the GARCH and multiple regression models, they found that the exchange rate does indeed have an effect on movement of the gold and oil prices and stock market prices. Zhou et al. [48] used the daily stock market (high frequency) data for the year 2016 and the Generative Adversarial Networks model using the Long Short-Term Memory (LSTM) and Convolutional Neural Networks, to predict the stock market prices in China.

Apart from these conventional economic determinants of the stock market performance, the sentiments and expectations have been used to predict stock indices. Bouktif et al. [5] have examined the use of sentiment features and provided an intricate pipeline to extract sentiment from tweets in the context of stock market. They used stock prices for 10 companies from NASDAQ from the year 2008 to 2018. A comparison between various models such as Logistic regression, Random Forest and Neural Networks is shown and the study highlighted how sentiment features help to increase the accuracy of the model. The complexity of the stock market signal is handled in Jin et al. [21] by using the Empirical Modal Decomposition method before feeding the features to the LSTM model which uses its memory units to optimize for timeseries data. Their work shows that the model also decreases the training time as well as increases prediction accuracy. The Stock tweets platform is used to scrape text and extract sentiment for the period from 2013 to 2018.

Furthermore, Twitter was used as the data source to monitor the public reaction and health during disasters e.g. hurricanes [22,37,47,48], floods [18], earthquakes [15], terrorist bombing Buntain et al. [9] public health related misinformation propagation [8,29,40] and others [19,42,44] and disease outbreaks [14,26,28,41].

Against this backdrop, our primary objectives in this paper are – (i) to study if Twitter sentiment and healthcare data during the pandemic periods are useful factors to predict stock market index movements, (ii) to compare the performance of different models for the prediction, and (iii) to study how the regression performance differs with respect to different sectors in United States (US) and India. The data used in this study are mainly of 4 kinds: Daily economic indicators, COVID-19 Health, Twitter Sentiment, Stock Index/ Price Related. Since our period of study is constrained to only 5 months, in order to have sufficient data points we use economic indicators that are available on a daily basis. This restricted us from using many other important indicators which do not have a daily frequency.

The results show that in this case, LSTM model shows better performance when compared to linear regression and ARIMA which may be due to its ability to capture nonlinear relationships between the features especially with respect to sentiment. When the sentiment-based features are used, the prediction ability of the model is improved and hence the sentiment features may provide additional useful information to model the stock price. The sentiment relationship with index is stronger in case of Finance and Consumer Goods sectors indices which have in general been more affected due to the pandemic unlike the Technology and Healthcare sectors. The relationship is in general stronger in US than in India which can be explained by greater percentage of people who are active on Twitter. The model performs best when the time lag is 1 day that is if we use the twitter sentiment of the previous day to predict the current stock price movement.

Materials and methods

The empirical analysis is carried out following the work-flow schematic presented in Fig. 1.

Data

The data is mainly collected for the period January - May 2020 (daily) as this was when the pandemic spread throughout the world and its effects on the economy were clearly evident.
Healthcare Data: The healthcare data was collected from [46] which is known to be a standard and accurate source of COVID-19 healthcare data. Apart from the global data, the dataset was grouped according to the country. The main variables that we use/ derive for our study are - daily number of active cases, number of deaths and number of recovered cases. We also calculate the corresponding daily rates for each of the mentioned variables. The testing coverage data is used to normalize the variables so as to level out the disparity in the number of cases which may be due to a low testing coverage in the country.

Economic: The Indian economy indicators used are mainly collected from the RBI’s daily press releases on Money Market Operations, Weekly Statistical Supplement [31] and India Government Bonds India Government Bonds, [20] which provides the daily Bond Yield Rates. The corresponding data for the US are collected from [32]. Foreign exchange rate is one of the variables considered as India has a floating exchange rate which depends on the demand and supply and is a good indicator of the economy.

Apart from this, the other indicators that are used are mainly related to the Money Market. The central bank of most countries maintains the liquidity which is the cash available in the economy by using monetary policies. The various measures taken, affect the way in which Banks respond with changes in interest rates among others. It also helps manage the amount of inflation and activity in the economy. It has short term effects on financial markets. Repo combines lending-borrowing and sale-purchase transactions. The securities are sold in exchange for cash and these securities can then be repurchased by the original holder. The reverse repo is used to reduce or absorb liquidity. Thus, we consider those variables that vary on a daily basis and therefore do not include important rates such as the Repo Rate, Reverse Repo Rate among others. It should be noted that The Net Liquidity injections also include other operations done to change liquidity but which may not be done on a regular basis. The central bank uses bond to borrow money. Government bonds (G-secs/ Treasury) are considered to be very safe investments. The yield rate changes with the price of the bond. The rise and fall of yields capture the expectation of stakeholders about the future growth of the economy. For example, the 10Y bond yield rate in the US has decreased owing to the worsening of the expectations of the economy due to COVID-19. Many of these factors are interdependent on each other by being only correlated or by having a causal relationship with a certain time lag. Since the time lag cannot be estimated directly we try different combinations and experiment to get the best results. The foreign exchange rates depend on various variables including the overall economic activity, prospects, etc. several factors a nation’s economic activity and growth prospects, interest rates, and geopolitical risk. Thus, the following variables are used: [Reverse Repo Operations (INR), Net Liquidity Injection (INR or $), Cash Balance (INR or $), 10Y Bond Yield Rates (%), 5Y Bond Yield Rates (%), 1Y Bond Yield Rates (%), Forex Rate ($/ INR)]

Stock Data: The data consists of the index values for major stock markets in the US and India collected from [33] and (NSE, [27]). The sector wise index (DJIA and Nifty) data is collected for the Financial, Consumer Goods, Healthcare and Technology sectors. Fig. 2 shows the timeseries data of Dow Jones U.S. Technology Index starting from 6 June 2019 to 6 June 2020.
Twitter Data: The twitter data was mainly collected from (PanaceaLab, [30]) and consisted only of tweet IDs. The tweets were categorized based on the location into Indian, US and others. The tweets were mainly filtered based on the keywords such as ‘COVID-19’, ‘Corona’, ‘Wuhan’, etc. The number of tweets is around a million.

Methods

Data pre-processing

Data pre-processing consists of various operations. Better predictive models can be built more efficient in terms of convergence and prediction using pre-processing operations. Some of the techniques that are commonly used with respect to numerical type data (since our data is numeric type) have been discussed below:

- Minmax scaling for numerical features using normalization, feature or sample wise (done by setting the axis of normalization). This is done by computing the overall minimum and maximum values of the feature taking all samples into consideration.
- Standard scaling (z-score normalization) numerical features using $\mu$ and $\sigma$ computed on the training dataset i.e. subtracting mean and dividing by standard deviation. It is also called Zero-mean scaling.
- Bucketizing features (of numeric type) using quantiles.
- Filling in missing values in the data using the median (for numerical features), mean or mode (for categorical features).
- Computing the PCA of the input features to project the data into a lower dimensional space (with linearly dependent features) to reduce the input size and prevent overfitting.

The min-max normalization procedure is used in this study to convert the data in the range between 0 and 1. We apply data pre-processing techniques on all the stock price of the indices that we collected for Indian and U.S. sectors. The DJIA and Nifty sector-wise indices are normalised using the min max scaling. After this operation, the stock prices obtained are in the range 0 to 1. This technique performs better if the distribution is not Gaussian or the standard deviation is very small. The sentiments obtained from the sentiment analyser are already scaled from $-1$ to $+1$, hence they do not require any scaling as such. The health data for number of COVID cases is also scaled before using in the model.

Sentiment analysis

Using the tweet ids based on the location, we use the Twitter Search API to fetch the tweet contents. The sentiment is based on 3 dimensions: Valence, Arousal and Subjectivity which are all values from $-1$ to $1$. By projecting the emotion along 3 different dimensions we can get useful information about the different aspects of the overall sentiment. Valence is a measure of how positive or negative the tweet is, Arousal is the intensity of the emotion and Subjectivity is used to ascertain if the text is an opinion or factual content. Each tweet must be pre-processed through an elaborate pipeline so as to get the best features and to remove noise. Some of the pre-processing includes:

- Removal of stop words which removes some of the common words such as ‘in’, ‘a’, etc. which do not add useful information.
- Named Entity and Parts of Speech tagging.
- Many tweets carry many slang words and contracted forms of words. We use a slang word dictionary to replace slang words and expressions to appropriate meaningful words.

![Fig. 2. Dow Jones U.S. Technology Index starting from 6 June 2019 to 6 June 2020.](image-url)
• The contracted words are expanded using a custom dictionary that maps the contracted versions of the word to the expanded version of the word(s).
• Correcting misspelled words and malformed sentences by replacing stray characters, URLs, username mentions, etc.
• We check if a particular tweet has a hashtag which corresponds to an emotional word and use this as well to predict sentiment.
• Sentences with lot of trailing shows high arousal thus this is taken into account.
• Emoticons are mapped to emotional states.

The Affective Norms for English Words (ANEW), consists of around 1000 words which are manually and accurately annotated a mean valence and arousal score, by experts [7]. The AFINN dataset, [16] also provides a similar list but with only Valence scores. These two lists were further extended by using the WordNet [25]. For every word in ANEW and AFINN, we add the same scores for all words in the synset of the word. This way we have a much larger list of EUs (Emotional Units). A valence and arousal score can be directly assigned to the words in the tweets that are present in the expanded ANEW and AFINN.

Principal component analysis

The Principal Component Analysis (PCA) is a linear method of dimensionality reduction that forms the underlying basis of multivariate data analysis based on projection methods. In PCA, the data is projected along the directions of in which there is maximum variance. Eigenvalues and eigenvectors of the covariance (see eqn. (1)) matrix of data are computed and each eigenvector corresponds to a principal component. The suitable number of features that can be extracted from this method is given by the number of the non-null eigenvalues or the number of principal components that can explain a statistically significant amount of variance in the data, generally kept as 99% explained variance. The functional form of PCA is –

\[ \text{Cov}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y}) \]

(1)

This method gives us a projection of the features (or high dimensional data points) in a vector space of a lower dimension. The principal components are mutually orthogonal. Fig. 3 shows the principal components for a dataset in 2 dimensions or a dataset with two features. In this work, we employ PCA to reduce the number of features to a few so that the correlated features can be minimized and only essential regressors can be considered for the prediction or classification model.

Auto regressive integrated moving average (ARIMA) model

Time Series forecasting can be of tremendous commercial value. Businesses use it for predicting demand and sales to manage the procurement of resources. The regression function can be written as a variable defined as a linear combination of the others -

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n \]

(2)

Regression can be linear or non-linear, depending on the variables and their correlation with other considered features. When considering a time series type of data, regression is used with respect to forecasting. It can be primarily of two types:
• Uni-variate Time Series Forecasting: The previous values of the time series alone are used to predict its future values.
• Multi-variate Time Series Forecasting: Predictors other than the past values of the series (also called as exogenous variables) are used for forecasting.

For our analysis, we use the Auto Regressive Integrated Moving Average (ARIMA) model and the linear regression model, two of the most used regression models under the univariate and multivariate timeseries data respectively.

An ARIMA, is a type of forecasting algorithm model used for statistical analysis. It uses time-series data to decipher patterns so as to predict future trends. While seasonal or periodic data can be best described by exponential smoothing
models, ARIMA models describe the autocorrelations in the data. Here, we limit our experimentation to only non-seasonal ARIMA models. To describe the model, we segregate the terms in it and study each term individually.

Autoregression model: Forecasting a variable using a regression of past values of the variable as regressors, hence the term “auto” regressors. Thus, an autoregressive model of order p can be written as:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \epsilon_t,$$

where, $\epsilon_t$ is white noise and $y_{t-i}$ denotes the i-th backward lag of $y_t$. We refer to this as an AR(p) model, an autoregressive model of order p, also called as the lag order.

Moving average models: It uses past forecast errors in a regression-like fashion. It is referred to as an MA(q) model of order q, wherein q refers to the number of lag error values used.

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q},$$

Furthermore, stationarity is a desired property in timeseries data for it to be used in forecasting. A time series is called stationary if its properties do not depend on the time at which the series is observed. Often, the timeseries data in level are not stationary. Hence to convert them into stationary series, the raw observations are differenced. The symbol $d$ is used to show the number of times that the raw data are differenced. It is also known as the degree of difference ($d$).

Giving the input parameters as p, q and d, we can estimate the regression coefficients and assess if past values of stock do indeed have an effect on the present value. Since past literature on the study of stocks with respect to this model suggests short term causations, we used $d$ as 1, i.e., single-differencing of raw data and q as 0, i.e., no past error is used. However, the p-value depends on the stock being assessed. So, we estimated the p-value for each stock such that the stock behaves as stationary. Apart from the AR and MA terms, other independent variables are included in the ARIMA model estimated in this study.

Long-Short-Term-Memory (LSTM) networks

Traditional neural networks cannot learn from past events and are not designed to persist information to use as a formulation for understanding current or future events. This seems like a major shortcoming. Recurrent neural networks (RNN) address this issue as they are networked with loops in them, allowing information to persist. A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor.

However, sometimes, we only need to look at recent information to perform the present task instead of the total past events. This issue of long-term dependencies is addressed by Long-Short-Term-Memory networks (LSTM’s). LSTM can help us model non-linear dependencies. This is done by using a series of “gates” (Fig. 4). These gates are contained in similar memory blocks which are connected layer by layer:

An LSTM unit consists of three types of gates: Input Gate: Scales input to cell (write) Output Gate: Scales output to cell (read) Forget Gate: Scales old cell value (reset). Each gate can be thought of as a switch that controls the read or write, incorporating the long-term memory function into the model (Fig. 4).

The present study applies several inputs like twitter sentiments, economic indicators, healthcare indicators and stock market indicators to LSTM. Iterative LSTM parameter tuning was performed for data fitting. Several hyperparameters were evaluated to identify the optimal LSTM architecture that provides ideal assessment metrics. The hyperparameters include:

1. number of LSTM layers;
2. number of nodes in each layer;
3. number of fully connected layers;
4. types of activation function;
5. number of dropout layers and percentage of dropout;
6. learning rate;
7. loss function;
8. optimizer;
9. batch size; and
10. number of epochs

The LSTM structure for the study comprises the following layers –
1. LSTM with 512 nodes.
2. Fully connected layers with 64 nodes and ReLU activation function
3. Dropout with 0.4%.
4. Fully connected layers with 1 node and Linear activation function.

The final hyperparameters are:
1. the learning rate is 0.001;
2. loss function of mean absolute error (MAE);
3. optimizer: ADAM;
4. epochs is 500; and
5. the batch size is 8

Though LSTM’s and ARIMA might look the same in principle, both have their pros and cons. The advantages of using ARIMA over LSTM’s are:
• Simple to implement, no parameter tuning
• Easier to handle multivariate data
• Quick to run

However, LSTM’s have proven to show better results more often and are hence more popular in time-series modeling for the following reasons:
• No pre-requisites (stationarity, no level shifts)
• Can model non-linear function with neural networks
• Needs more data

Results and discussion

The sentiment analyser is able to accurately capture the trends which vary according to the healthcare events that dominate every user’s Twitter feeds. In Fig. 5, it can be seen that on the day the WHO declared COVID-19 as an international concern and a health emergency, there is a sudden dip in the emotion i.e. there is a sudden increase in negative sentiment (Fig. 6).
The following is a sample run of the sentiment analyser.

Valence: $-0.31$
Arousal: $+0.44$
Subjectivity: $0.78$

The analyser is able to accurately capture the negative sentiment (negative valence), the intensity of the emotion (high arousal) and the opinionated nature of the text (high subjectivity).

Next, we perform correlation analysis by computing the Pearson correlation coefficients between the different regressors. The value of the coefficient primarily captures the variation along the line of best fit. We use 16 samples to observe any trend between the features. From the analysis, we aim to determine if there exists a linear or direct correlation between the stock prices and the twitter sentiment. We get a value of $0.634$ as the correlation coefficient, which might not be as too significant for small data size. It is possible that the same value of $0.6$ would be highly significant when the size of the data is huge. However, since our sample size is small, we cannot derive any strong inference about the association between the sentiments and the stock prices and further analysis is needed.

It is also possible that the features might be non-linearly related and since the Pearson correlation can only capture linear association, it is not able to produce a statistically significant result. This indicates that LSTM might have an advantage over linear techniques like ARIMA and LR in predicting the stock, if there is any sort of correlation between the features (sentiment and stock prices) at all.

We discuss three potential models with respect to stock price prediction. However, for our analysis, it would be difficult to interpret and manage results from all these models. Hence, we decide to derive further results by choosing the most suitable model for our data. This was implemented by running a sample analysis and comparing the Root Mean-Squared Error from all models. At last, we choose the model with the least RMSE. Using all the predictors as input, we fit all the models for finding the model with best fit.

For ARIMA, we try with values of $p$ from 0 to 6 and derive the best fit. Finally, we use the $p$, $q$, and $d$ values as 1, 1 and 0 respectively. Standard values of $q$ and $d$ are 1 and 0 respectively. For ARIMA, we use the stats model python library and multivariate ARIMA where we pass the other features as a numply array argument to the function separately (exogenous variables).

For LSTM, its multiple input series method is used where all the features are given as inputs to form a 2D array, derived at the same time step (i.e.1.day). A simple LSTM with the standard architecture is used without many LSTM layers. The used architecture is shown in the Fig. 4. The optimizer used is Adam and the loss function as Mean Squared Error (MSE). It is run for a total of 30 epochs as the model can be seen to converge by 10 epochs. The linear regression is run for simple OLS estimate.

As can be seen from Table 1, the LSTM model has the highest accuracy in predicting both Indian and U.S. stock index values of the technology companies, as this model has the lowest RMSE value compared to both linear regression and ARIMA models. This is in conjunction with the past studies that prove the credibility of LSTM for its higher stock price prediction ability. Linear regression model performs worse than ARIMA, possibly because the ARIMA can predict based on previously observed values of the time series (though linear) and is able to map more complicated relations. The success of LSTM and ARIMA over LR depict that there is indeed some dependence of the stock price on its previous values. From here on, we show all results using the LSTM model.

Further, to derive whether sentiments do have some correlation with the stock price, we fit the model with all the features and one without sentiment features. Tables 2 and 3 report the RMSE values from prediction of stock index values of finance and technology sectors respectively. We find that RMSE increases when the sentiment-based features are not used. This helps us to conclude that the twitter sentiments do have some impact on the stock price during this pandemic. This can be supported by the fact that due to social distancing more and more people expressed their opinions through social media platforms which is in turn affecting the sentiment of investors more profoundly. This effect is more evident in U.S. than in India, possibly because U.S. has a higher number of people expressing themselves over social media. This has

### Table 1

RMSE for different models.

|          | ARIMA | Linear Regression | LSTM |
|----------|-------|-------------------|------|
| US-Tech  | 1.047 | 1.248             | 0.874|
| India-Tech | 1.153 | 1.290             | 0.536|

Note: RMSE values are of order $10^2$, & MSE of the order $10^6$. 

![Sample Tweet](Image)
been corroborated with two industries, namely Finance and Technology (Tables 2 and 3). We infer that the Tech industry shows lesser improvement in the RMSE than the Finance. This may indicate that the Finance industry is likely more affected in the pandemic and is more correlated to public’s current sentiments.

Next, we performed the PCA to reduce the feature size to a manageable count. Initially, the feature set size is 16. After observing the features closely, 3 broad categories of features are found, namely healthcare, sentiment and macroeconomic indicators. Therefore, we perform PCA to reduce the feature set to 3. To support our assertion, the RMSE values of the prediction models before and after applying PCA are derived. As expected, the RMSE increases as the principal components are used instead of the original features as dimensionality reduction often leads to loss of desirable information. However, the increase in RMSE can be traded off for the lesser features.

We analyze how each sector in India and U.S. is affected by the pandemic, and more importantly how the twitter sentiments impact the same. As shown in Table 4, in the Indian scenario, the stock indices of the Consumer goods and Finance sectors are much more determined by the COVID-19 impact as compared to the stock indices of the Pharma and Technology sectors, with consumer goods being the most affected. The RMSE value of the LSTM prediction model is lowest for the Consumer goods sector followed by Finance sector. Similarly, the $R^2$ value, which is used as a measure of goodness of fit of a model, is highest in case of Consumer goods sector followed by the Finance sector. In the U.S scenario, as depicted by Table 5, the Financial Sector is most impacted by COVID-19 sentiments. Furthermore, we examined with what time lag the Twitter sentiment must be used to achieve most accurate prediction and found out that a lag of 1 day in Twitter sentiment is most useful in predicting the current stock market movement. Thus, this study sheds light on several interesting aspects of movement of the economic market during the COVID-19 pandemic. Although the results here do not derive a strong relationship between the factors, they provide further evidence and show scope for providing useful information to help predict economic movement which can further be taken up in subsequent studies.

### Table 2
RMSE of LSTM model for Finance Sector in India and U.S.

|                     | India- Finance | U.S.- Finance |
|---------------------|----------------|---------------|
| Without Sentiment features | 0.5142         | 0.5109        |
| With Sentiment features      | 0.4017         | 0.3010        |
| Principal components        | 0.4229         | 0.3119        |

Note: RMSE values are of order $10^3$, & MSE of the order $10^5$.

### Table 3
RMSE of LSTM model For Technology Sector in India and U.S.

|                     | Indian-Tech | U.S. Tech |
|---------------------|-------------|-----------|
| Without Sentiment features | 0.9102      | 0.6917    |
| With Sentiment features      | 0.8018      | 0.5075    |
| Principal components        | 0.8744      | 0.5364    |

Note: RMSE values are of order $10^3$, & MSE of the order $10^5$.

### Table 4
$R^2$, RMSE and MSE Metrics of LSTM model for Indian Sectors.

|        | $R^2$  | RMSE  | MSE  |
|--------|--------|-------|------|
| Financial | 0.5647 | 0.4229 | 0.1788 |
| Pharma   | 0.5315 | 0.5654 | 0.3196 |
| Tech     | 0.5019 | 0.8744 | 0.7645 |
| Consumer Goods | 0.5783 | 0.3723 | 0.1386 |

Note: RMSE values are of order $10^3$, & MSE of the order $10^5$.

### Table 5
$R^2$, RMSE and MSE Metrics of LSTM model for U.S. Sectors.

|        | $R^2$  | RMSE  | MSE  |
|--------|--------|-------|------|
| Financial | 0.5710 | 0.3119 | 0.0972 |
| Pharma   | 0.5475 | 0.4634 | 0.2147 |
| Tech     | 0.5142 | 0.5364 | 0.2877 |
| Consumer Goods | 0.5333 | 0.4923 | 0.2423 |

Note: RMSE values are of order $10^3$, & MSE of the order $10^5$. 

---

P.R. Jena and R. Majhi

Scientific African 19 (2023) e01480
Conclusion

In this study, we evaluated the usefulness of the Twitter sentiment for improving the prediction of stock market indices of several sectors in the U.S. and India. We described different methods involved in extracting the sentiment, pre-processing and the techniques applied for using these models to predict the stock market indices of four sectors. Mainly, the recurrent neural network model, namely the LSTM and the conventional linear timeseries models such as ARIMA are compared for their prediction accuracy. The LSTM model is found to be more accurate across the sectors. The variation of the results across different sectors has been examined and it is found that Finance and Consumer Goods sectors have a stronger relationship with the health and Twitter sentiment features more than the Technology and Pharmaceutical sectors. This may be due to the fact that these sectors are more adversely affected due to the pandemic. Especially, the Consumer Goods sector felt a huge demand side shock as consumers across the globe either experienced a complete lockdown and hence, not allowed to visit shops and malls, or were subjected to movement restriction that resulted in a drastic reduction in demand. Supply chain disruptions due to the pandemic severely affected the movement of consumer goods from production to distribution centres further affecting the consumer sentiment. Similarly, the Finance sector was also affected by the negative sentiments of the investors who felt safe to hold on to their money rather than investing in different financial instruments. On the other hand, since the Pharmaceutical and Technology sectors were mostly seen as the ones that developed products that can help people overcome the pandemic effects. Hence, the negative sentiments for these sectors were significantly less. Furthermore, the relationship between twitter sentiments and stock market performance was stronger in the case of US which denotes the higher percentage of people in that country who actively engage with the Twitter platform.

The contribution of the current study is it establishes a significant link between non-monetary factor such as social media sentiments and stock market variations. Though, some of the past studies have examined this link, they mostly used the linear timeseries models. We developed a nonlinear neural network model that could use the nonlinearity in the data and predicts with higher accuracy. Future work will mainly involve studying more sectors and using stronger economic indicators. Moreover, the interpretability of the models can be improved by trying other neural network models.

Data availability statement information

The data that support the findings of this study are available from the corresponding author upon request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Atsalakis, K. Valavanis, Surveying stock market forecasting techniques - Part II: Soft computing methods, Expert Syst. Appl. 36 (3 PART 2) (2009) 5932–5941, doi:10.1016/j.eswa.2008.07.006.
[2] R. Aguilar-Rivera, M. Valenzuela-Rendón, J.J. Rodríguez-Ortiz, Genetic algorithms and Darwinian approaches in financial applications: a survey, Expert Syst. Appl. 42 (21) (2015) 7684–7697.
[3] B. Barber, O. Terrance, All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors, Rev. Financ. Stud. 21 (2008) 785–818.
[4] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[5] S. Bouktif, A. Fiaz, M. Awad, Augmented textual features-based stock market prediction, IEEE Access 8 (2020) 40269–40282, doi:10.1109/ACCESS.2020.2976725.
[6] M.A. Boyacioglu, D. Avci, An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange, Expert Syst. Appl. 37 (12) (2010) 7908–7912.
[7] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEW): instruction manual and affective ratings, IEEE Internet Comput. 12 (5) (2008) 44–52, doi:10.1109/MIC.2008.114.
[8] D.A. Broniatowski, A.M. Jamison, S.H. Qi, L. AlKulaib, T. Chen, A. Benton, S.C. Quinn, M. Dredze, Weaponized health communication: twitter bots and Russian trolls amplify the vaccine debate, Am. J. Public Health 108 (10) (2018) 1378–1384, doi:10.2105/AJPH.2018.304567.
[9] C. Buntain, J. Golbeck, B. Liu, G. LaFree, Evaluating public response to the boston marathon bombing and other acts of terrorism through Twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, 10, 2021, pp. 555–558, doi:10.1609/icws.m.v10i11.14774.
[10] O. Bustos, A. Pomares-Quimbaya, Stock market movement forecast: a systematic review, Expert Syst. Appl. 156 (2020) 113464.
[11] R. Cavalcante, R. Brasileiro, V.Souza, J. Nobrega, A. Oliveira, Computational intelligence and financial markets: a survey and future directions, Expert Syst. Appl. 55 (2016) 194–211, doi:10.1016/j.eswa.2016.02.006.
[12] Q. Chen, I. Goldstein, W. Jiang, Price informativeness and investment sensitivity to stock price, Rev. Financ. Stud. 20 (3) (2007), doi:10.1093/rfs/hjl024.
[13] N.Y. Choi, R.W. Sias, Why does financial strength forecast stock returns? Evidence from subsequent demand by institutional investors, Rev. Financ. Stud. 25 (2012) 1550–1587.
[14] M. Dredze, D.A. Broniatowski, K.M. Hillyard, Zika vaccine misconceptions: a social media analysis, Vaccine 34 (30) (2016) 3441–3442, doi:10.1016/j. vaccine.2016.05.008.
[15] P.S. Earle, D.C. Bowden, M. Guy, Twitter earthquake detection: earthquake monitoring in a social world, Ann. Geophys. 54 (6) (2011) 708–715, doi:10.4401/ag-5364.
[16] Finnrup Nielsen, 'Making Sense of Micro posts': big things come in small packages, in: CEURWorkshop Proceedings, 2011, pp. 93–98.
[17] H. Harapan, N. Itoh, A. Yufika, W. Winardi, S. Keam, H. Te, D. Megawati, Z. Hayati, A.L. Wagner, M. Mudatsir, Coronavirus disease 2019 (COVID-19): a literature review, J. Infect Public Health 13 (5) (2020) 667–673, doi:10.1016/j.jiph.2020.03.015.
[18] E. Hirata, M.A. Giannotti, A.P.C. Larocco, J.A. Quintanilha, Flooding and inundation collaborative mapping – use of the Crowdmap/Ushahidi platform in the city of Sao Paulo, Brazil, J. Flood Risk Manag. 11 (2018) 598–5109, doi:10.1111/jfr.12181.
[19] M. Imran, P. Mitra, C. Castillo, Twitter as a lifeline: human-annotated Twitter corpora for NLP of crisis-related messages, in: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC, 2016, 2016, pp. 1638–1643.

[20] India Government Bonds - Investing.com India. (n.d.). Retrieved July 31, 2020, from https://in.investing.com/rates-bonds/india-government-bonds

[21] Z. Jin, Y. Yang, Y. Liu, Stock closing price prediction based on sentiment analysis and LSTM, Neural Comput. Appl. 32 (13) (2020) 9713–9729, doi:10.1007/s00521-019-04504-2.

[22] M.Y. Kabir, S. Madria, CoronaVis: A Real-time COVID-19 Tweets Analyzer, 2020 http://arxiv.org/abs/2004.13932.

[23] Li, Y. (2012). Empirical study on the relationship between money supply and stock market in Europe. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7473 LNCs, 539–544. https://doi.org/10.1007/978-3-642-34062-8_70

[24] N. MECHRI, S. Ben Hamad, C. de Peretti, S. Charfi, The impact of the exchange rate volatilities on stock markets dynamics: evidence from Tunisia and Turkey, SBN Electron. doi:10.2139/ssrn.3396400.

[25] G.A. Miller, WordNet: a lexical database for English. Commun. ACM 38 (11) (1995) 39–41, doi:10.1145/219717.219748.

[26] R. Nargar, Q. Yuan, C.C. Freifeld, M. Santillana, A. Nojima, R. Chunara, J.S. Brownstein, A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives, J. Med. Internet Res. 16 (10) (2014) e236, doi:10.2196/jmir.3416.

[27] NSE - National Stock Exchange of India Ltd. (n.d.). Retrieved July 31, 2020, from https://www.lseindia.com/

[28] M. Odlum, S. Yoon, What can we learn about the Ebola outbreak from tweets? Am. J. Infect. Control 43 (6) (2015) 563–571, doi:10.1016/j.ajic.2015.02.023.

[29] S.O. Oyeyemi, E. Gabarron, R. Wynn, Ebola, Twitter, and misinformation: a dangerous combination? BMJ 349 (October) (2014) 14–15, doi:10.1136/bmj.g6178.

[30] PanaceaLab - COVID19 Twitter Dataset Homepage. (n.d.). Retrieved July 31, 2020, from http://www.panaceaalab.com/covid19/

[31] Reserve Bank of India - Database. (n.d.). Retrieved July 31, 2020, from https://www.rbi.org.in/Scripts/Statistics.aspx

[32] Resource Center. (n.d.). Retrieved July 31, 2020, from https://www.treasury.gov/resource-center/Pages/default.aspx

[33] S&P Dow Jones Indices. (n.d.). Retrieved July 31, 2020, from https://www.spglobal.com/spdji/en/

[34] T.A. Saleh, Protocols for synthesis of nanomaterials, polymers, and green materials as adsorbents for water treatment technologies, Environ. Technol. Innov. 24 (2021) 101821.

[35] T.A. Saleh, Nanomaterials and hybrid nanocomposites for CO2 capture and utilization: environmental and energy sustainability, RSC Adv. 12 (37) (2022) 23860–23888.

[36] T.A. Saleh, Global trends in technologies and nanomaterials for removal of sulfur organic compounds: clean energy and green environment, J. Mol. Liq. (2022) 119340.

[37] A. Sebastian, W. Mobley, W. Highfield, S.D. Brody, Leveraging Machine Learning and Twitter Data to Identify High Hazard Areas during Hurricane Harvey, c. 2019, p. 12549.

[38] M. Sirucek, Munich Personal RePeC Archive Macroeconomic Variables and Stock Market : US Review Macroeconomic Variables and Stock Market : US review, 39094, 2012.

[39] C. Sohrabi, Z. Alsaif, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iossifidis, R. Agha, World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19), Int. J. Surg. 76 (February) (2020) 71–76, doi:10.1016/j.ijsu.2020.02.034.

[40] B.G. Southwell, J. Niederdeppe, J.N. Cappella, A. Gaywensky, D.E. Kelley, A. Oh, E.B. Peterson, W.-Y.S. Chou, Misinformation as a Misunderstood Challenge to Public Health, Am. J. Prev. Med. 57 (2) (2019) 282–285, doi:10.1016/j.amepre.2019.03.009.

[41] Szomszor, M., Kostkova, P., & De Quincey, E. (2011). #Swinfluenza: twitter predicts swine flu outbreak in 2009. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 69 LNICST (December 2009), 18–26. https://doi.org/10.1007/978-3-642-43358-3_3

[42] Z. Wang, N.S.N. Lam, N. Obradovich, X. Ye, Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data, Appl. Geogr. 108 (April) (2019) 1–8, doi:10.1016/j.apgeog.2019.05.001.

[43] E. WEISBROD, Stockholders' unrealized returns and the market reaction to financial disclosures, J. Finance LXIV (2) (2019) 899–942.

[44] D. Wladdimiro, P. Gonzalez-Cantergiani, N. Hidalgo, E. Rosas, Disaster management platform to support real-time analytics, in: Proceedings of the 2016 3rd International Conference on Information and Communication Technologies for Disaster Management, ICT-DM 2016, 2017, doi:10.1109/ICT-DM.2016.7857207.

[45] World Economic Outlook UpdateA Crisis Like No Other, An Uncertain Recovery, June 2020.

[46] WorldometerCoronavirus CASES, Worldometer, 2020, doi:10.1101/2020.01.23.20018549V2.

[47] J. Yang, M. Yu, H. Qin, M. Lu, C. Yang, A Twitter data credibility framework—Hurricane Harvey as a use case, ISPRS Int. J. Geoinf. 8 (3) (2019), doi:10.3390/jgi8030111.

[48] X. Zhou, Z. Pan, G. Hu, S. Tang, C. Zhao, Stock market prediction on high-frequency data using generative adversarial nets, Math. Probl. Eng. (2018), doi:10.1155/2018/4907423.