Social Network Analysis and Mining (2022) 12:92
https://doi.org/10.1007/s13278-022-00919-3

ORIGINAL ARTICLE

Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak

Nabanita Das[1] · Bikash Sadhukhan[1] · Tanusree Chatterjee[1] · Satyajit Chakrabarti[2]

Received: 1 March 2022 / Revised: 27 June 2022 / Accepted: 4 July 2022 / Published online: 27 July 2022

Abstract

Forecasting the stock market is one of the most difficult undertakings in the financial industry due to its complex, volatile, noisy, and nonparametric character. However, as computer science advances, an intelligent model can help investors and analysts minimize investment risk. Public opinion on social media and other online portals is an important factor in stock market predictions. The COVID-19 pandemic stimulates online activities since individuals are compelled to remain at home, bringing about a massive quantity of public opinion and emotion. This research focuses on stock market movement prediction with public sentiments using the long short-term memory network (LSTM) during the COVID-19 flare-up. Here, seven different sentiment analysis tools, VADER, logistic regression, Loughran–McDonald, Henry, TextBlob, Linear SVC, and Stanford, are used for sentiment analysis on web scraped data from four online sources: stock-related articles headlines, tweets, financial news from "Economic Times" and Facebook comments. Predictions are made utilizing both feeling scores and authentic stock information for every one of the 28 opinion measures processed. An accuracy of 98.11% is achieved by using linear SVC to calculate sentiment ratings from Facebook comments. Thereafter, the four estimated sentiment scores from each of the seven instruments are integrated with stock data in a step-by-step fashion to determine the overall influence on the stock market. When all four sentiment scores are paired with stock data, the forecast accuracy for five out of seven tools is at its most noteworthy, with linear SVC computed scores assisting stock data to arrive at its most elevated accuracy of 98.32%.

Keywords Web scraping · Sentiment analysis · Stock market prediction · Deep learning

1 Introduction

Stock market forecasting continues to be a challenging task in the economics sector due to its extremely stochastic character. Forecasting and analysing stock market movements have acquired huge notoriety, as stock market movement changes may have a profound influence on the economy. Political, social, environmental, economic, and public health factors all have an impact on stock market movement (Chou, Park, and Chou, 2021; Shang et al. 2021), causing markets to oscillate and become complex and uncertain (Chaudhuri, Mukherjee, Chowdhury, Sadhukhan, and Goswami, 2018; Wagner 2020). The stock market's volatility is well known to investors. They constantly monitor market movements to manage micro-investments and maximize profits while minimizing risk. Predicting stock market movement is a difficult task that requires much data analysis. Appropriate statistical models and artificially intelligent algorithms are required to address these issues and find an adequate solution. Numerous machine learning and deep learning algorithms may produce a reliable forecast with minimal errors (Mukherjee, Sadhukhan, Sarkar, Roy, and De, 2021).

Stock market movement can be studied using fundamental analysis (which considers economic considerations) or technical analysis (which considers historical data) (Valle-Cruz et al. 2021). Investors' opinions, traders' feelings, general public views, and different news items are another category of factors that undoubtedly influence the stock market (Biswas et al. 2020). It may collectively be classified as part of the well-known field of research known as sentiment analysis. Sentiment analysis is a type of analysis that uses statistics, natural language processing, and machine learning to ascertain the emotional content of communications (Hajhmida and Oueslati 2021; Hussein 2018).

✉ Bikash Sadhukhan
bikash.sadhukhan@tict.edu.in

[1] Department of Computer Science & Engineering, Techno International New Town, Kolkata, West Bengal, India

[2] University of Engineering and Management, Kolkata, West Bengal, India

COVID-19 was found for the first time in India in January 2020. It could have caused a terrible pandemic. Since March 2020, all workplaces, including offices, shops, and markets, have been shut down indefinitely. All commercial activities were halted, resulting in economic collapses around the world. People are forced to work from home due to the total lockdown scenario. During this hard time, social media platforms are profoundly used to share feelings, opinions regarding economic issues, and the dilemma in stock market investments. Opinions and feelings are posted on many social media platforms, and financial news and articles are in several languages from various Indian states. Natural language processing assists in their processing, and sentiment analysis extracts their feelings (Rajput 2020).

Sentiment analysis can be led through an assortment of approaches and tools. Sentiment analysis is currently receiving much attention for predicting stock market movements. This study focuses on the sentiment analysis of tweets, Facebook comments, news headlines, and online financial news articles. The emotion ratings generated in this manner are paired with stock data to investigate the repercussions of a COVID-19 pandemic. The motivation behind this exploration is to introduce a model in which sentiment scores produced by multiple sentiment analysis techniques are integrated with stock market data to quantify and compare the prediction performances. Seven sentiment analysis tools are utilized in this article to construct sentiment scores from four different sources of web scraped data. The data for the Nifty-50 stock market index were obtained from Yahoo Finance for this research. Stock data have been used to extract OHLC (open, high, low, and close) characteristics.

The rest of this research work is organized as follows: Section 2 discusses related works done in this field of research. Section 3 describes the background studies involved in this work. Section 4 presents the main system model proposed in this research work. Section 5 illustrates the experimental analysis and implementation. Section 6 discusses the results and their analysis. Section 7 compares the proposed work with existing works. Finally, Sect. 8 precisely concludes the work with some future work proposals.

2 Related work

The recent rise in the availability of textual data has prompted a surge in interest in sentiment analysis. Opinion mining and opinion summarizing are the two main subfields of sentiment analysis. The former is often concerned with forecasting whether the text reflects a positive or negative value based on what we are attempting to predict, whereas the latter is typically concerned with summarizing what has been stated (Derakhshan and Beigy 2019). Sentiment analysis may be performed at various levels of abstraction. This section focuses on in-depth reviews of various relevant research articles. The primary focus in this case is to examine stock market movement prediction and sentiment analysis of web scraped data.

Numerous researchers have collected and analysed Facebook comments to use them in various operations and decision-making processes (Akter and Aziz 2016; Hajhmida and Oueslati 2021; Marengo et al. 2021; Rase 2020). Hajhmida et al. proposed using Facebook data for the prediction of mobile application breakout. They used the Facebook graph API to evaluate the sentiment polarity of user comments and then built a breakout prediction model using machine learning techniques (Hajhmida and Oueslati 2021). Akter et al. established market prices by employing sentiment analysis of data acquired from FOODBANK's social media posts, which is a very popular Facebook group in Bangladesh, using the lexicon approach (Akter and Aziz 2016). Marengo et al. used a language modelling approach to explore connections between language stated on Facebook and self-reported quality of life (physical, psychological, social) (Marengo et al. 2021). Deep learning technologies such as convolutional neural networks and long short-term memory have been utilized to understand people's feelings and opinions by producing sentiment analysis of Afaan Oromoo social networking site information such as Facebook posts and comments (Rase 2020).

Twitter sentiment analysis also enables us to make numerous decisions. They utilized an LSTM model that includes investor feelings, stock price time series data, and an attention mechanism to provide an accurate forecast of stock prices (Chou et al. 2021). Investors' emotions are taken into account, and tweets from investors are collected and sorted using a sentiment index to determine whether the investor plans to purchase or sell. Hassan et al. analysed the sentiments stated in tweets about new research publications to assess how influential they are early in the research cycle. According to the findings, a positive association between tweet emotions and citation counts was shown to be useful in predicting the early impact of literature (O. A.-H. Hassan, Ramaswamy, and Miller, 2009). Lu et al. performed sentiment analysis on a large dataset of tweets related to cruise tourism during the COVID-19 pandemic.

The study highlights the significance of sentiment analysis and reaffirms a recent request for sentiment analysis to be a critical component of tourism research (Lu and Zheng 2021). Public sentiment may be connected with stock price behaviour. Kordonis et al. used machine learning techniques to determine the correlation between tweets and stock market price behaviour (Kordonis, Symeonidis, and Arampatzis, 2016). Forecasting election results also makes use of sentiment analysis, which analyses public opinion on social media to make accurate predictions about how voters will support (Chauhan et al. 2021).

Newspaper articles and headlines are another source of text for sentiment analysis. Ghasiya et al. used the non-negative matrix factorization (NMF) topic modelling technique on Middle East-related articles from three Japanese newspapers. After the identification of critical themes, they employed typical supervised machine learning techniques to extract overall and topic-specific sentiments from the acquired headlines (Ghasiya and Okamura 2021). Users' sentiments obtained from news headlines have a significant impact on traders' buying and selling behaviours, since they are quickly influenced by what they read. Gite et al. utilized LSTM-based deep learning in conjunction with machine learning techniques to anticipate stock prices with a high degree of accuracy (Gite et al. 2021). Mehta et al. developed and deployed a technique for predicting the accuracy of stock prices that takes public opinion into account in addition to other characteristics. To estimate future stock prices, the suggested algorithm takes into account public sentiment, opinions, news, and past stock prices (Mehta et al. 2021).

Online financial news and other news articles are crucial tools for making many decisions, which may be used in a variety of research areas through sentiment analysis. A novel sentiment analysis system based on a deep neural network was developed in (Shi et al. 2021). The novel technique improved sentiment categorization by 9% when compared to the logistic regression method. Additionally, the sentiment information calculated by the analysis system was applied to the stock movement prediction job and significantly enhanced performance when compared to techniques that used simply trading data as input. Ly and Nguyen aimed to mitigate investor risk by developing a revolutionary framework that uses sentiment analysis to anticipate the first three, five, ten, twenty, and thirty days of an IPO's price movement by evaluating its prospectus (Ly and Nguyen 2020). Wu et al. calculated the investors' sentiment index using a sentiment analysis approach based on convolutional neural networks using nontraditional data. They integrated sentiment index, technical indicators, and historical stock transaction data as the stock price prediction feature set and used a long short-term memory network to forecast the China Shanghai A-share market (Wu et al. 2021). When forecasting the daily price trend of the OMXS30 stock market index, researchers found that adding sentiment characteristics extracted from financial news to a numerical dataset based on past prices improved classification performance (Elena 2021). Arif et al. examined the performance of learning classifier systems (LCSs), which are rule-based machine learning approaches, in sentiment analysis of tweets and movie reviews, as well as spam identification using SMS and email datasets. (Arif et al. 2018). The existing LCS approach is expanded by incorporating a unique encoding scheme for classifier rules to account for feature vector sparsity. The collected findings indicate that the suggested encoding strategy accelerated the learning process and consistently produced high-quality outcomes across all studies. Turner et al. emphasized stock price prediction using a sentiment vocabulary constructed from financial conference call records. They provided a technique for automatically generating an emotion lexicon based on an established probabilistic methodology. The research further demonstrates that when forecasting stock price change, domain-specific sentiment lexicons outperform general sentiment lexicons (Turner, Labille, Computer Science and Computer Engineering, University of Arkansas, Fayetteville, Arkansas, United States, Gauch, and Computer Science and Computer Engineering, University of Arkansas, Fayetteville, Arkansas, United States, 2021). Huang and Tanaka designed a modularized multiagent reinforcement learning system with the goal of introducing scalability, reusability, and depth of information intake to financial portfolio management using web news sentiment data (Z. Huang and Tanaka 2021). They demonstrated that their technique qualifies as a stepping stone for inspiring further innovative financial portfolio management system designs by its originality and superiority over current benchmarks. Another recent study aims to forecast the erratic price movement of cryptocurrencies by studying social media sentiment and determining their association (X. Huang et al. 2021). The research presented a method for determining the sentiment of messages on China's most popular social media network, Sina Weibo. In this research, Weibo posts were captured, the crypto-specific sentiment lexicon was created, and a long short-term memory (LSTM)-based recurrent neural network was used to forecast the price trend for future time frames using the past cryptocurrency price movement. Table 1 shows a brief summary of related work in this domain.

According to the review study, significant research has previously been done on sentiment analysis in stock market movement prediction using web scraped data from various sources, such as Twitter, Facebook, and news headlines. However, significant additional work is required to correctly estimate the influence of public sentiment on stock market movement.

3 Methodologies

This section precisely covers the main theoretical notions used in the current research endeavour. Sentiment analysis is the well-known approach of data science. Almost every active research area employs data science approaches in their respective areas, since it brings together many algorithms, machine learning theories, and tools to unearth buried knowledge from raw data (Budiharto 2021). Currently, stock market movement prediction and analysis is one of the most popular domains where data science is used

Table 1 Summary of related work

Web scraping source	Related work	Source data particulars	Sentiment analysis (SA) tools and other working algorithms	Result
Facebook	(Hajhmida and Oueslati 2021)	Web crawling apptracs.com. Comments, No. of Shares, Likes and other relevant information stored in MongoDB	Lexicon of user created words, nearest neighbours, RBF SVM, decision tree, random forest, Neural Nets, AdaBoost, Naïve Bayes, logistic regression	Polarity of highly positive, positive, highly negative and negative. Best prediction accuracy 84.73% from random forest
	(Akter and Aziz 2016)	Bangladesh-based Facebook group "FOODBANK"	Dictionary-based; Lexicon-based, Naïve Bayes	Dictionary-based accuracy 73%, Lexicon-based analysis outperforms Naïve Bayes
	(Marengo et al. 2021)	603 user-generated languages	Data mining through LIWC closed vocabulary method utilizing Random Forests to predict QoL dimensions	Highest accuracy achieved in Psychological and general QoL dimensions
	(Rase 2020)	1452 comments of Oromo Democratic Party's official site	Document-level sentiment by multinomial Naïve Bayes, LSTM and CNN	Although MNB outperforms both LSTM (87.6% accuracy) and CNN (89% accuracy), it faces problems of indirect comments
Twitter	(Chou et al. 2021)	Stock Twits and Twitter for SA, Yahoo Finance for Stock data	LSTM and GloVec	LSTM + Sentiment Score + attention Model outperforms
	(S.-U. Hassan et al. 2020)	Tweets for SA and Altmetric.com for publications	SentiStrength, linear regression with one additional indicator: 'number of unique Twitter users'	Positive correlation between Tweets and literature's early impact
	(Lu and Zheng 2021)	Tweets	LDA model, Kullback–Leibler (KL) divergence	Eagerness for niche cruise rather than mass cruise
	(Mehta et al. 2021)	Apache Flume used for Tweets of Bitcoin, News articles	XGBoost, LSTM	Positive sentiment towards prediction of cryptocurrency
	(Singh et al. 2021)	tweepy APIs from 20 Jan 2020 to 25 April 2020 for world tweet data and Indian's tweets	BERT tool for classification, VADER for intensity, and TextBlob for polarity and subjectivity	Indians communicated positively towards Govt. activity with 94% accuracy
	[18]	100,000 Tweets using API duration 19/10/2020–29/10/2020	TextBlob Rule Based, LDA, SARIMAX	76% accuracy, RMSE 0.196
	(Chauhan et al. 2021)	Tweets	Machine learning, Lexicon-based and Deep learning	Application of machine learning methods dominated in election result prediction
News paper	(Ghasiya and Okamura 2021)	Headlines of three Japanese newspapers	Topic modelling approach NMF, NLP and ML Algorithms	50.37% negative and 49.63% positive result
	(Gite et al. 2021)	"The Pulse" for SA and Yahoo finance for Stock Data	LSTM-CNN for SA and LSTM for stock price prediction	93.15% accuracy
	(Mehta et al. 2021)	BSE Sensex-Infosys for Stock data and Money control, IIFL, Economic Times, Business Standard, Reuters, and Live Mintdata for SA	Support Vector Machine, MNB classifier, linear regression, Naïve Bayes and Long Short-Term Memory	LSTM outperforms with 92.45% accuracy

Table 1 (continued)

Web scraping source	Related work	Source data particulars	Sentiment analysis (SA) tools and other working algorithms	Result
Online financial news and other news articles	(Shi et al. 2021)	Snowball financial online community in China for collecting financial comments of investors for SA, The Shanghai exchanges, the top 50 stocks in Shenzhen exchanges and the top 30 stocks in American stocks	CNN, GRU for SA. SVM, LR for stock prediction	9% improvement over LR for SA, Stock prediction is improved 1.25% over LR
	(Ly and Nguyen 2020)	Five different datasets: first 3, 5, 10, 20, and 30-days price	EDGAR package and the Loughran–McDonald Sentiment Word Lists for SA, Baseline model, Random Forests, Decision Tree, Naïve Bayes and Logistic Regression for price movement forecasting	Logistic Regression performs best, then comes Naïve Bayes and Baseline model
	(Wu et al. 2021)	Stock posts and financial news for SA and China Shanghai A-share market data	CNN for SA and LSTM for Stock Closing Price prediction	Accuracy is very close to the actual price
	(Elena 2021)	OMXS30 stock data and financial news for SA	Vader, Loughran–McDonald for SA and a tree-based ensemble model: XGBoost for Stock Price Movement Prediction	A hyper parameter is extracted using cross-validation and grid search for better performance
	(Arif et al. 2018)	Web data and Kaggle dataset	Naïve Bayes, RCNN and Random Forest	Kaggle dataset outperform by 96.13% over 86.5% from web-scrapped data. Naïve Bayes performs least efficiently and RCNN most efficiently
	(Turner et al. 2021)	Stock data and web data with domain-specific lexicon	Henry's lexicon, Loughran's lexicon and SentiWordNet	Domain-specific lexicon is most accurate
	(Z. Huang and Tanaka 2021)	Historical prices from US stock markets and asset related news from media	Deep Q-Network	EAM-enabled SAM performed best
	(X. Huang et al. 2021)	Weibo social media posts for SA and Crypto currency price	LSTM	LSTM performs better than the AR approach by 18.5% in precision and 15.4% in recall

extensively. The movement of market prices is impacted by a variety of online factors, including social media comments, financial news, stock-related news, and many more. Natural language processing is a method of dealing with these sorts of unstructured online data by turning them into a structured format that a computer can combine with stock data to determine their influence on market prediction (Biswas et al. 2020; Hajhmida and Oueslati 2021; Hussein 2018).

3.1 Natural language processing, sentiment analysis, and web scraping

Natural language processing (NLP) is a subfield of artificial intelligence (AI) that analyses text data to uncover underlying knowledge (Okon et al. 2020). Stock market fluctuations may have a significant economic impact on the economy and individual customers. On the other hand, public events, inflation, and the news media all have an effect on stock price movement (Shi et al. 2021). Sentiment analysis is a classification technique that addresses public data for opinion mining. It employs natural language processing techniques to determine the polarity of an opinion, emotion, or feeling in terms of positive, negative, or neutral sentiments (Elena 2021). Web data are necessary in a wide variety of fields, including research, academia, business, marketing, and governance. These data are available in a variety of formats. Manually downloading web data is a tedious task. Web scraping is a data extraction technology offered as a software application that automatically extracts data from different websites and stores it in a common type of database, allowing for easier processing, analysis, and visualization of data (De S Sirisuriya, 2015; Patel 2020).

3.2 Sentiment analysis tools

Each of the seven sentiment analysis tools utilized in this work is depicted below. The locations of these instruments in the sentiment analysis categorization are depicted in Fig. 1. Although logistic regression (LR) is classified as supervised learning regression, it conducts binary classification. As a result, logistic regression and support vector classifiers are two of the most extensively used classification techniques. Logistic regression is a statistical technique that utilizes a linear dataset to predict binary values for any number of independent variables (Ly and Nguyen 2020). On the other hand, SVC generates an optimal separating hyperplane to

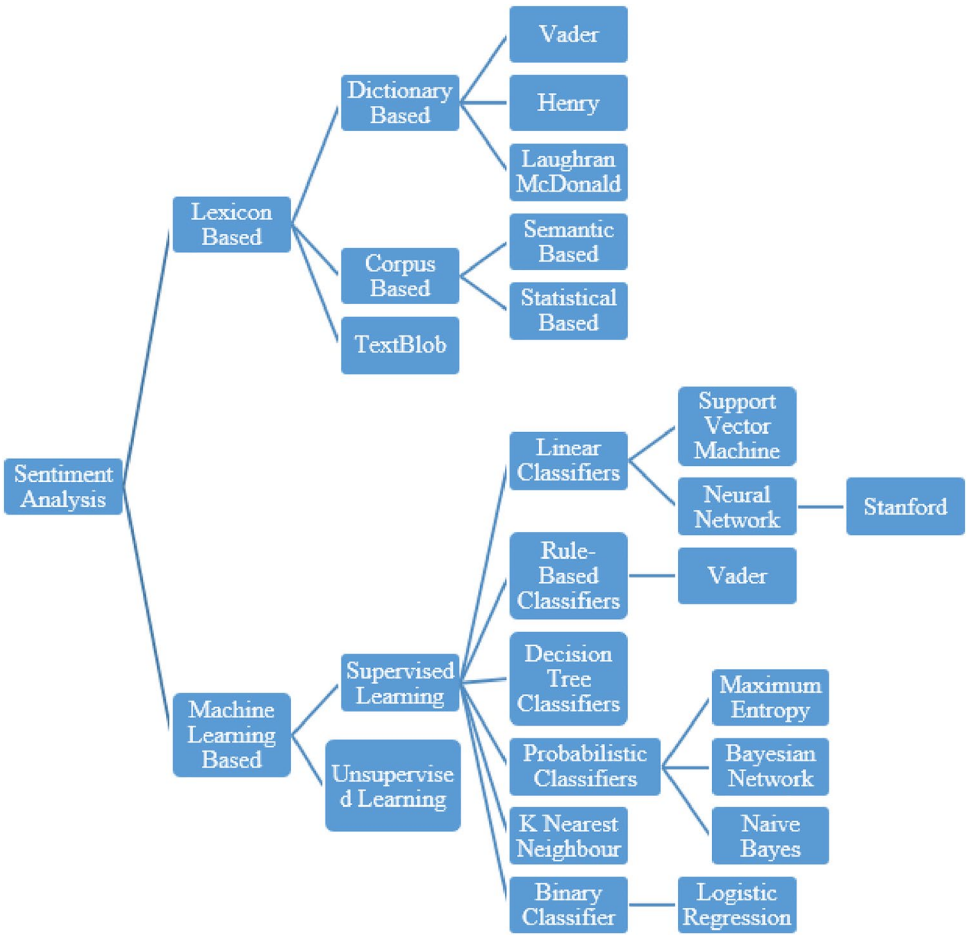


Fig. 1 Sentiment analysis tools' (used in this work) positions in sentiment analysis classification

discover maximum data separation during classification (Mehta et al. 2021).

NLTK (Natural Language Tool Kit) is a premier open-source natural language processing (NLP) platform for Python, featuring over 50 corpora. SentiWordNet is a lexicon resource and text processing tool for classification, tokenization, and semantic reasoning. TextBlob is a Python library that reuses NLTK corpora to assign polarity and subjectivity scores to the text data after processing (Bonta et al. 2019). The Valence Aware Dictionary for Sentiment Reasoning, abbreviated VADER, is a lexicon-based and rule-based free, open-source sentiment analysis tool that classifies polarity (positive, negative, neutral) as well as the degree of polarity value word by word (Singh et al. 2021). It works better with social media data and uses the polarity_scores () function to calculate the polarity of words (Bonta et al. 2019). The Loughran–McDonald tool is equipped with a sentiment dictionary with six sentiment dimensions based on the financial industry, best-suited for financial text classification (Elena 2021). Positive, negative, and neutral polarity classifications are generated using this manually created dictionary. If C-1 and C-2 denote the positive and negative word counts, respectively, then C-1/sentence and C-2/sentence are used to denote the sentence's polarity as positive (1) or negative (-1)(Turner et al. 2021). Henry is another dictionary-based sentiment analyser. As with Loughran–McDonald, this tool is focused on positive and negative words associated with finance. The loadDictionaryHE () function is used to access the words during this analysis process (Turner et al. 2021). Stanford CoreNLP is a sentiment analysis tool that is based on a recursive neural network. It computes the sentiment score as polarity by examining the meaning of the text (Lin et al. 2018).

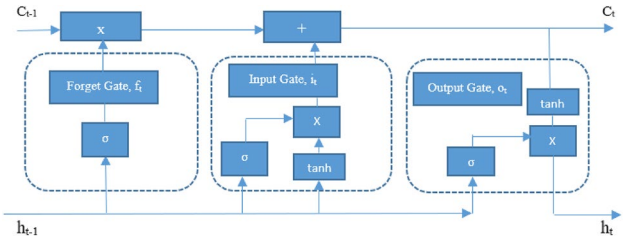


Fig. 2 LSTM architecture

3.3 Long short-term memory (LSTM)

Stock market data are time-series data that can be processed and used efficiently by LSTM, an improved version of RNN to forecast future price movements (Mehta et al. 2021). Figure 2 shows the comprehensive LSTM architecture (Van Houdt et al. 2020). With three gates, an input gate, an output gate, and a forget gate, LSTM overcomes RNN's inability to remember long-term dependencies by preserving relevant information and erasing no relevant information (Gers and Schmidhuber 2001). The forget gate preserves relevant long-term data using Eq. (1), the input gate updates information using σ as the excitation function as in Eq. (2), and finally, the output gate provides output with Eq. (3). Equation (4) generates output vector h_t.

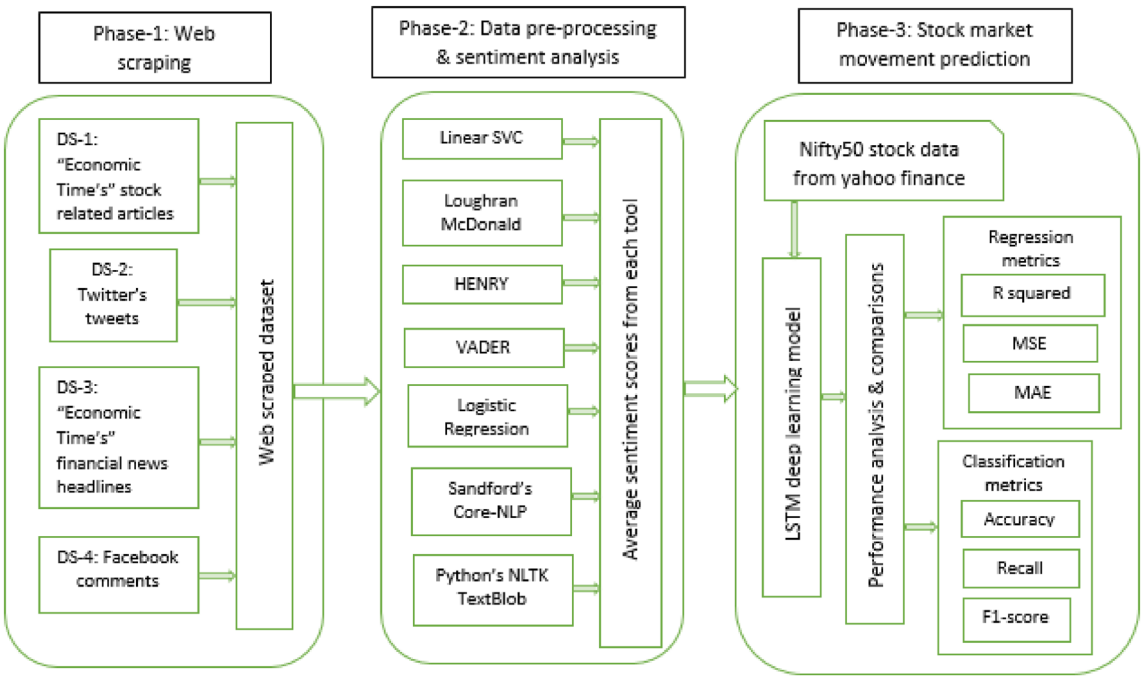


Fig. 3 System model of the current research work

$$f_t = \sigma\left(w_{xf}x_t + w_{hf}h_{t-1} + w_{cf}c_t + b_f\right) \quad (1)$$

$$i_t = \sigma\left(w_{xi}x_t + w_{hi}h_{t-1} + w_{ci}c_t + b_i\right) \quad (2)$$

$$o_t = \sigma\left(w_{xo}x_t + w_{ho}h_{t-1} + w_{co}c_t + b_o\right) \quad (3)$$

$$h_t = o_t \tanh\left(c_t\right) \quad (4)$$

where σ is the activation function, w is the weight matrix, b_f, b_i, and b_o are deviation vectors, x_t is the input vector, c_{t-1} is the old cell state, c_t is the updated cell state, and h_t is the output vector.

4 Proposed model

This section describes in detail the proposed model and algorithms used in this research work by a systematic three-phase working structure.

4.1 System model

Figure 3 displays the system model of this research, which is a three-phase architectural model.

The entire processing is done through three phases:

Phase-1: This phase is in charge of scraping public opinions, online news, and articles from related web pages. There were many scraping sources in this phase, such as social media, online news articles, and financial news headlines.

Phase-2: This phase carries out data preprocessing and sentiment analysis of the scraped data done in Phase-1. Disparate sentiment analysis tools calculate sentiment scores after performing preprocessing on the scraped data fed in from Phase 1. Preprocessing is a vital step before sentiment analysis to obtain better accuracy.

Phase 3: Phase 3 is the final phase of the proposed system model. This phase accumulates the calculated sentiment scores performed in Phase-2 with the stock data to calculate stock market movement prediction with the help of the LSTM deep learning model. The experiment thus performed is classified into regression and classification modes. In both classifications of experiments, two categories of results are generated for analysis. One is the data-tool combination to produce the best stock market movement prediction performance, and the other is determining stock market movement prediction performance with the combined effect of sentiment scores. The calculated results are analysed and compared to come up with different conclusions.

4.2 Algorithm

In this subsection, three algorithms for each phase are described. Algorithm 1 depicts the web scraping mechanism performed in Phase-1, Algorithm 2 illustrates the sentiment analysis in Phase-2, and Algorithm 2.1 outlines preprocessing, which is the fundamental task before performing sentiment analysis. Finally, Algorithm 3 does the final trick in this research project: It predicts how the stock market will move based on the sentiment scores that were calculated in Phase 2.

4.2.1 Algorithm 1: web scraping

Web scraping is performed for a specific duration from 'N' different sources to gather 'N' sets of raw text data used in sentiment analysis in Phase-2. Table 3 provides the implementation details of the web scraping in this work in Sect. 5.

Algorithm 1: Web-scraping

Input: Tokens of different sources// *Twitter's Tweets, Facebook's Comments, Economic Times's (ET's) Headlines and articles*

Output: Web-scraped data DS-1 through DS-N stored in.csv file for N number of sources// *Tweets, Comments, Headlines and Articles*

Step-1: Select number of sources, N

Step-2: Choose platform// *Python for Twitter, Facebook and ET' Headlines, JAVA for ET's articles*

Step-3: Set up scraping platform by importing required packages specific for each source

Step-4: For *i=1* to N,

    Apply API access tokens for Source[*i*] and set them up in the platform

Step-5: End For

Step-6: Choose keyword and/or topic for scraping sources// *Keyword for Twitter and Facebook, topic for ET*

Step-7: Determine time *range//Date range*

Step-8: Apply scraping with calling specific library functions

4.2.2 Algorithm 2: data preprocessing and sentiment analysis

'N' sets of web scraped raw data are required to be preprocessed before feeding into the 'M' number of sentiment analysis tools. Data need to be preprocessed to achieve higher accuracy while performing sentiment analysis.

The data preprocessing function of the sentiment analysis algorithm is illustrated in Algorithm 2.1 next to Algorithm 2. A sentiment score of 1 is assigned for positive sentiment, -1 for negative sentiment and 0 for neutral sentiment. Then, the daywise average sentiment score is calculated. Each set of data (Set-1 to Set-N) contains the daytime average sentiment score calculated from each sentiment analysis tool (Tool-1 to Tool-M). Table 4 depicts the data format used in this study prior to and during the execution of the sentiment analysis algorithms.

Algorithm 2: Data preprocessing and sentiment analysis

Input: Datasets DS-1 through DS-N stored in.csv file// *Twitter's Tweets, Facebook's Comments, Economic Times's (ET's) Headlines and articles*

Output: Average sentiment scores per day in.csv file// *1 for positive, -1 for negative and 0 for neutral*

Step-1: For i=1 to N,
Preprocessing()// Preprocessing function call
Step-2: End For
Step-3: Choose p*latform// Python for Tweets, Facebook Comments and ET' Headlines, JAVA for ET's articles*
Step-4: Select M number of tools, Tool[1] through Tool[M]// *Logistic Regression, VADER, SVC, etc.*
Step-5: Set up platform by importing required packages specific for each Tool
Step-6: Choose preprocessed training dataset to train the model
Step-7: Apply vectorization
Step-8: Evaluate model
Step-9: For i=1 to M
Perform testing with different datasets, DS-1 through DS-N with Tool[i]// *Tweets, Comments, Headlines and Articles*
Step-10: End For
Step-11: For i=1 to N
Step-12: For j=1 to M,
Calculate average sentiment scores per day as ADS-[i]_T[j]
Step-13: End For

Algorithm 2.1: Preprocessing()

Input: Datasets DS-1 through DS-N stored in.csv file// *Twitter's Tweets, Facebook's Comments, Economic Times's (ET's) Headlines and articles*

Output: Preprocessed web scraped raw datasets

Step-1: For i=1 to N,
Preprocess each web scraped datasets DS[i]
Step-2: Preprocess with the following operations:
Step-2.1: Removing stop-words, punctuations, extra whitespaces, extra newline and duplicate entries
Step-2.2: Appropriate packages should be imported for removing links, usernames and *emojis// Package "re" is used to replace the links, usernames and emojis*
Step-2.3: Breaking down sentences into tokens
Step-2.4: *Lemmatizing// Reduce certain words to their basic form*
Step-2.5: Converting all tense to present t*ense// Only present tense is allowed*
Step-2.6: Converting all the words into lowercase
Step-3: End For

4.2.3 Algorithm 3: stock market movement prediction using LSTM model

The Phase-3 algorithm will combine the daytime average sentiment scores produced in phase-2 ('N' X 'M') with Nifty50 stock data collected from Yahoo Finance and then put them into the LSTM-based stock market movement prediction process to improve performance. This phase will analyse and compare the influence of 'N' X 'M' sentiment scores on stock market movement forecasts in terms of the regression metrics MSE, MAE and R-squared and the classification metrics accuracy, recall and F1 score. Table 5 specifies the implementation details.

Algorithm 3: Stock market movement prediction using LSTM model

Input: Average sentiment scores calculated in Algorithm 2 from different tools in.csv file and stock *data// ADS-[i]_T[j]*

Output: Performance in terms of accuracy in percentage

Step-1: Download stock data for specific time-range
Step-2: Set up platform by importing required packages
Step-3: Prepare training dataset to train the model
Step-4: Prepare testing dataset by combining average sentiment scores calculated in Algorithm 2 with stock data:
Step-5: For i=1 to M,*// M=Number of sentiment analysis tools*
Step-6: For j=1 to N,*// N= Number of datasets from M sources*
Stock data + ADS-[j]_T[i]
Step-7: End For
Step-8: End For
Step-9: Preprocess dataset by normalization,$z = \frac{x - \min(x)}{\max(x) - \min(x)}$
Step-10: Construct LSTM model by setting all its parameters
Step-11: Train and test LSTM with the datasets to generate performance in terms of accuracy in percentage

Table 2 Details of the experimental set-up

Machine set-up	Programming platform and corresponding tools	Sentiment analysis tools	Data sources: Time range: 1 July 2020 to 31 December 2020
Windows 10	Anaconda3	Logistic Regression, Linear Support Vector Classifier,	Nifty50 Stock Data from Yahoo Finance
Intel Core i5	Jupytar Notebook 5.7.8	Vader, Stanford's Core-NLP, Textblob	Web scraped from Facebook, Twitter,
8 GB RAM	Keras 2.2.4, JAVA 8	Henry, Loughran-McDonald	"Economic Times" Stock Headlines and Financial News Article from "Economic Times"

Table 3 Details of web scraping implementation

Data	Sources of web scraping 01/07/2020 to 29/12/2020	Web scraping methods	Scraped raw data size
DS-1	Stock related articles headlines from Economic Times	Selenium and pandas	1266
DS-2	Tweets from Twitter with keyword "nifty50"	Twint rather than Tweep which can only extract tweets upto the last 7	79,908
DS-3	Financial news from Economic Times	Eclipse, Jdk 8 or above, Maven, Selenium framework (findElements(By.tagname()), Chrome Webdriver	295
DS-4	Facebook comments with keywords like nifty finance, nifty stocks, nifty prediction, nifty analysis, nifty advice, nifty trend, nifty 50	Facebook being a dynamically loaded website, Python's ever popular library called "beautifulsoup" can't be used for web crawling. Instead, a web automation tool called "selenium" has been used for this purpose	341

5 Experimental analysis and implementation

This section focuses on the experimental set-up and implementation process to achieve the intended outcome.

5.1 Experimental set-up

Table 2 portrays the details of the experimental set-up to implement and execute the obligatory experiments in this research work.

5.2 Implementation

This subsection illustrates the implementation process of the three phases of the proposed system model, as depicted in Fig. 3.

5.2.1 Web scraping

Web scraping is the first phase of the system model depicted in Fig. 3. In this study, four sources of online data ('$N'=4$ in Algorithm 1) are scraped to feed into phase 2 of the system model for preprocessing and sentiment analysis. Table 3 provides a description of the web scraping that has been performed along with the scraped raw data size. Scraped data are stored in a csv file. Data from four different online sources are shown in Table 8. Scraped raw data are represented as DS-1 through DS-4, where $N=4$ (number of online data sources).

5.2.2 Data preprocessing and sentiment analysis

The second phase of the system model is devoted to sentiment analysis. Web scraped raw data from Phase-1 need to be preprocessed before performing sentiment analysis. Four independent sets of scraped raw data (DS-1, DS-2, DS-3, and DS-4) are preprocessed using the methods described in Algorithm 2.1 in this study.

The preprocessed data in the.csv file are then supplied into seven distinct sentiment analysis tools ('$M'=7$ in Algorithm 2). The Seven Tools are mentioned in Table 2. These seven tools analyse the sentiment of each of the four data points, and each tool generates a daily sentiment score for each opinion, news or article gathered per day from each of the four data points. Therefore, a total of 28 ('N x $M'=4\times7$) distinct sentiment ratings were created for use in Phase-3 of the system model. The implementation of the seven tools is discussed as follows:

5.2.2.1 Logistic regression and linear SVC: Training and testing data in csv files have been preprocessed and are ready to be put into the logistic regression model/linear SVC model. Pickle, seaborn, nltk, sklearn, matplotlib, and a number of additional packages must be imported. Here, Twitter sentiment analysis is mentioned. The model is trained using a training.csv file containing 79,908 tweets. Tweets, their feelings, date, user, flag, and ID are all included in the file. Training requires only text and feelings. The processed text vectorization is then performed. The pickle package may be used to save the models in a pickle file. The test data are now turned into a list that the LR model/linear SVC model can read. The findings are saved in a separate.csv file with two columns: tweets and feelings. Each tweet has its own sentiment.

5.2.2.2 Stanford's core NLP: Stanford's Core-NLP is needed to insert NLP requisite open-source libraries of JAVA language to add the dependency in the pom.xml file. Documents are iterated by passing preprocessed data (DS-1 through DS-4) one by one into a document. In the next step, the sentiment method is used to obtain the sentiment scores and return them, which are saved in the.xlxs file. The Excel file has negative, positive, and neutral scores for the dataset.

Table 4 Details of data representation after web scraping and sentiment analysis implementation

Data	After web scraping	After sentiment analysis, average sentiment scores per day
Data-1	DS-1	ADS-1
Data-2	DS-2	ADS-2
Data-3	DS-3	ADS-3
Data-4	DS-4	ADS-4

Table 5 Details of data representation after toolwise sentiment analysis implementation

VADER	Logistic regression	Loughran–McDonald	Henry	TextBlob	Linear SVC	Stanford
ADS-1_V	ADS-1_LR	ADS-1_LM	ADS-1_H	ADS-1_TB	ADS-1_SVC	ADS-1_STF
ADS-2_V	ADS-2_LR	ADS-2_LM	ADS-2_H	ADS-2_TB	ADS-2_SVC	ADS-2_STF
ADS-3_V	ADS-3_LR	ADS-3_LM	ADS-3_H	ADS-3_TB	ADS-3_SVC	ADS-3_STF
ADS-4_V	ADS-4_LR	ADS-4_LM	ADS-4_H	ADS-4_TB	ADS-4_SVC	ADS-4_STF

Confusion Matrix		Actual	
		Positive	Negative
Prediction	Positive	TP	TN
	Negative	FP	FN

Fig. 4 Confusion matrix

5.2.2.3 Word Loughran–McDonald sentiment and henry sentiment To evaluate whether a statement is positive or negative, both approaches make use of dictionaries to classify words as positive or negative and then count how many times each positive or negative word appears in a given phrase. If the number of positive words exceeds the number of negative words, the statement is positive. The statement is negative if the number of negative words exceeds the number of positive words; otherwise, it is neutral. A data frame is created by assembling all of the scraped data (DS-1 to DS-4 one by one) line by line. Each word in each line is verified, and the positive counter is incremented if the word is positive. The negative counter is incremented if the term is negative. The counters were then compared to determine whether the line under experiment was positive, negative, or neutral. The sole distinction between these two dictionaries is the classification of terms as positive or negative.

5.2.2.4 VADER The SentimentIntensityAnalyzer package is first imported from the nltk.sentiment.vader module and then initialized. All scraped data (DS-1 to DS-4 sequentially) were examined line by line to calculate sentiment scores (positive, negative, neutral, and compound) and saved in a data frame where they scored the percentage of positive, negative, and neutral, which were then normalized to produce the compound score or "Overall Sentiment." The polarity of the compound score indicates the polarity of the line's "Overall Sentiment."

If the compound score is between −0.05 and 0.05, it is considered neutral; if the compound score is more than 0.05, it is considered positive; otherwise, it is considered negative. By using the above logic and comparing it to previous results, it was found that while the percentage of negative feelings remained constant, the percentage of positive sentiments grew. As a consequence of normalization for the purpose of calculating the compound score, the percentage of neutral feelings was reduced.

5.2.2.5 TextBlob Python has a library named "textblob" that is used to perform sentiment analysis by providing an interface for performing basic natural language processing activities. Each data entry is assigned a float polarity score of -1.0 for negativity and 1.0 for positivity by TextBlob. A score of 0 is awarded to circumstances in which no words map to any of the words in the pre-set training set.

Table 4 provides the symbolization of the data after performing web scraping and sentiment analysis. Table 5 shows the data representation following sentiment analysis using seven different sentiment analysis methods. Before sentiment analysis, the data are represented as DS-1 through DS-4, and after sentiment analysis, they are represented as ADS-1 through ADS-4, as average sentiment scores were calculated per day.

5.2.3 Stock market movement prediction using LSTM model

In Phase-3 of the proposed model, the tool-specific average sentiment scores derived in Phase-2 are sequentially integrated with Nifty50 stock market data acquired from Yahoo Finance. Each combination is entered into the LSTM model, which is used to forecast stock market movement. Accuracy of prediction is expressed as a percentage.

Many machine learning evaluation metrics have been used to estimate the performance of the proposed model (Batra and Daudpota 2018; Eck, Germani, Sharma, Seitz, and Ramdasi, 2021; Mokhtari et al. 2021). As the proposed model is implemented using LSTM, both regression and classification are implemented here to evaluate the model. Regression is evaluated in terms of R^2, MSE and MAE. The confusion matrix (Fig. 4) helped generate the accuracy, recall and F1 score of the classification implementation of the proposed model. The confusion matrix is a highly recommended metric to evaluate any machine learning prediction model in terms of TP, TN, FP and FN, where

1. TP (true positive): Correctly classified positivity.
2. TN (true negative): Correctly classified as negative.
3. FP (false positive): Falsely classified positivity.

Table 6 Details of Data Combination for input to the LSTM Model

Tool[i] generated sentiment scores combination with stock data for LSTM model				
Stock Data	Stock Data + ADS-1_Tool[j]	Stock Data + ADS-1_ Tool[j] + ADS-2_Tool[j]	Stock Data + ADS-1_ Tool[j] + ADS-2_ Tool[j] + ADS-3_Tool[j]	Stock Data + ADS-1_Tool[j] ADS-2_Tool[j] ADS-3_ Tool[j] + ADS-4_Tool[j]

Table 7 Details of the LSTM model specification

Model	Layers	Optimizer	Loss function	Classification metrics	Epochs
LSTM	3	Adam	MSE, MAE, R-squared	Accuracy, Recall, F1 score	100

4. FN (false negative): Falsely classified negativity.

The row containing TP and TN represents precision. The column consisting of TP and FP represents recall (sensitivity), whereas the second column consisting of TN and FN represents specificity. The metrics used in this research are recall, F1 score and accuracy. Equations (5), (6), and (7) are used to calculate the accuracy (correct percentage of prediction), recall (how many actual positives are predicted correctly) and F1 score (balance calculation between recall and precision) of the proposed model:

$$\text{Percentage Accuracy} = \frac{TP + TN}{TP + TP + FP + FN} * 100 \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - \text{score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}} \tag{7}$$

where Precision (how many positive predictions are correct) is expressed in Eq. (8).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

The pattern of data combinations used in this research is depicted in Table 6. The specifications for the LSTM Model are listed in Table 7. The mean square error (MSE) cost function is determined using Eq. (9):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\text{real}_i - \text{predict}_i\right)^2 \tag{9}$$

The mean absolute error (MAE) and R^2 are also calculated with Eqs. (10) and (11), respectively:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\text{real}_i - \text{predict}_i\right|\right) \tag{10}$$

$$R^2 = 1 - \frac{\sum_{1=1}^{n}\left(\text{predict}_i - \text{mean}\right)^2}{\sum_{1=1}^{n}\left(\text{real}_i - \text{mean}\right)^2} \tag{11}$$

In Eqs. (9), (10) and (11), `n' represents the number of samples.

6 Results and analysis

This section mainly describes the experimental results obtained during the research and its analysis.

6.1 Web scraping result

Web scraping of individual online data sources is displayed in Table 8 as a sample from every four sources.

6.2 Data preprocessing and sentiment analysis

Web scraped raw data from Phase-1 of the system model are passed into Phase-2 for raw data preprocessing and subsequent sentiment analysis. The following table shows the resultant sample of daily average sentiment scores obtained using three tools: VADER, linear SVC, and Henry. Table 9 depicts the average sentiment scores per day collected from different tools.

Table 8 Web-scraped data from four online sources: DS-1, DS-2, DS-3 and DS-4

Web scraping from four online sources		
Data	Date	Web-scraped data
DS-1: Stock Market Related Articles' Headlines:	30–12-2020	Trade Setup: Nifty prone to profit booking at current level, consolidation overdue
DS-2: Tweets from Twitter	31–12-202,023:57:00	The Nifty50 has finally hit the â,[1]14000 for the first time ever on the last day of 2020. I think the bull run is likely to continue in the year 2021, Nifty50 may hit 15,000 and Sensex to cross 50,000 by December #2021 #cryptocurrency #stock #indianstockmarket #intraday #india
DS-3: Financial News	Oct 28, 2020, 04:12 PM IST	Financial conditions in India have recovered significantly after hitting the abyss in April: Crisil
DS-4: Facebook Comments	15–12-2020	sensex inch fresh high hdfc twin sparkle bajaj finance top gainer sensex hdfc

Table 9 Average sentiment scores per day from different tools

Date	ADS-1_V	ADS-2_V	ADS-3_V	ADS-4_V
01–07-2020	0.75	0.737	1	1
02–07-2020	1	0.698	-0.25	0.5
03–07-2020	0.143	0.678	0.333	1
06–07-2020	0	0.713	1	1
07–07-2020	0.333	0.74	0.5	0
Date	ADS-1_SVC	ADS-2_SVC	ADS-3_SVC	ADS-4_SVC
01–07-2020	0.167	0.194	1	0.114
02–07-2020	0.111	0.132	0.333	0.129
03–07-2020	0.143	0.233	0.5	0.199
06–07-2020	0.5	0.217	1	0.177
07–07-2020	-0.167	0.204	0.286	0.183
Date	ADS-1_H	ADS-2_H	ADS-3_H	ADS-4_H
01–07-2020	0.583	0.192	0.139	0.66
02–07-2020	0.8	0.049	0.068	1
03–07-2020	0.078	0.069	0.078	0.74
06–07-2020	0.139	0.139	1	0
07–07-2020	0.154	0.106	0.333	0.154

Table 10 Sample Nifty50 stock data

Date	Open	High	Low	Close
07–01-2020	10,323.79	10,447.04	10,299.59	10,430.04
07–02-2020	10,493.04	10,598.20	10,485.54	10,551.70
07–03-2020	10,614.95	10,631.29	10,562.65	10,607.34
07–06-2020	10,723.84	10,811.40	10,695.09	10,763.65
07–07-2020	10,802.84	10,813.79	10,689.70	10,799.65

6.3 Stock market movement prediction using LSTM model

This section will mainly perform two categories of experiments, namely regression and classification, to determine the performance of stock market movement prediction in terms of the loss functions and percentage accuracy of the proposed model towards the prediction ability using the LSTM deep learning model. Sentiment scores calculated in Phase-2 are coupled with stock data in this phase. Table 10 contains some stock data as an illustration with graphical representation in Fig. 5.

Fig. 5 Stock price history

In this study, seven sentiment analysis tools are utilized to perform sentiment analysis on four web scraped data sources, yielding 28 sets of sentiment scores. Using the LSTM model, these sentiment ratings are paired with stock data to forecast market movement. The experimental findings were created and analysed into two categories.

6.3.1 Determining the data-tool combination to produce the best stock market movement prediction performance

Stock data are combined with each average sentiment score per day computed by seven different tools on four different scraped datasets to perform prediction operations pertaining to two experimental categories, regression and classification. Regression metrics are R-squared, MSE and MAE, and classification metrics are accuracy in percentage, recall and F1 score for inspecting tool level performances and data source level effectiveness of public sentiments on stock market movement. Table 11 displays the experimental results in terms of the regression metrics R-squared, MSE and MAE. Table 12 shows the experimental results in the form of classification metrics accuracy, recall, and F1 score. Table 11 presents stock data coupled with each sentiment score derived from seven tools on four data sources in columns and regression metrics as rows. From this table, we find that "Facebook Comments" (ADS-4) by linear SVC, Vader, and Loughran–McDonald yield the greatest results when paired with stock data independently. Next, better results come from sentiment scores, which were derived from financial news of "Economic Times" (ADS-3) from logistic regression and Henry. After that, tweets from Twitter (ADS-2)

Table 11 First category experimental results in terms of cost functions

Tool	Metric	Stock data with ADS-1	Stock data with ADS-2	Stock data with ADS-3	Stock data with ADS-4
Henry	R-squared	0.2098	0.1423	0.3523	0.2566
	MSE	0.0365	0.0441	0.0141	0.0263
	MAE	0.0929	0.0986	0.0855	0.0881
Logistic regression	R-squared	0.4015	0.1715	0.3237	0.2012
	MSE	0.0111	0.0422	0.0164	0.0316
	MAE	0.0795	0.0944	0.0859	0.0917
Loughran–McDonald	R-squared	0.2112	0.1689	0.1988	0.2714
	MSE	0.0301	0.0353	0.0321	0.0238
	MAE	0.0807	0.0934	0.0925	0.0853
VADER	R-squared	0.1864	0.2068	0.3037	0.3748
	MSE	0.0362	0.0326	0.0222	0.0133
	MAE	0.0986	0.09	0.0855	0.0823
TextBlob	R-squared	0.2282	0.2309	0.2008	0.2266
	MSE	0.0286	0.0282	0.0329	0.0299
	MAE	0.0912	0.0893	0.092	0.0913
Linear SVC	R-squared	0.3823	0.3311	0.3266	0.4176
	MSE	0.0115	0.0126	0.0133	0.0108
	MAE	0.0809	0.0829	0.0848	0.0781
Stanford	R-squared	0.1731	0.3201	0.1857	0.1666
	MSE	0.0398	0.0152	0.0382	0.0425
	MAE	0.0964	0.0863	0.0947	0.0951

perform best from TextBlob and Stanford. Finally, stock-related articles headlines from “Economic Times” (ADS-1) from logistic regression performed best. Among these best performing data-tool combinations in each group of experiments (prediction with sentiment scores from each data source and tool), it is clear that sentiment scores of "Facebook Comments," analysed by linear SVC when combined with stock data, generate the best accuracy measure of 98.11%.

Table 12 goes with the flow and portrays that VADER, linear SVC and Loughran–McDonald perform their best with ADS-4 with accuracy, logistic regression and Henry with ADS-3, TextBlob and Stanford with ADS-2 and finally logistic regression with ADS-1. Among these results, the linear SVC-generated sentiment score from ADS-4 has the best accuracy. Then, comes logistic regression from ADS-1 and VADER from ADS-4.

6.3.2 Determining stock market movement prediction performance with the combined effect of sentiment scores

Stock market movement prediction is performed by combining stock data with sentiment scores calculated from each of the seven tools from four sources one-after-another and checking the amalgamate effect on stock market movement prediction. As in the first category of experiments, Table 13 portrays the experimental results in terms of the regression metrics R-squared, MSE and MAE, and Table 14 shows the experimental results in the form of the classification metrics accuracy, recall and F1 score. When gradually combining each of the four sentiment scores with the stock data, five out of seven experiments show a significant increase in accuracy percentages. Figure 6 shows the performance results derived from each tool for stock data with combined datasets, i.e., ADS-1, ADS-2, ADS-3 and ADS-4. This figure shows the performance comparison of the effectiveness of four sentiment scores from four sources calculated by seven tools on stock market movement prediction.

All four linear SVC sentiment scores have a significant impact on stock market movement prediction, according to Fig. 6. In this case, the best prediction performance in percentage accuracy is 98.32 per cent for linear SVC, followed by 97.67 per cent for logistic regression and 96.85 per cent for VADER. Since linear SVC performs best, some of its experimental details are given in Table 15. Table 15 depicts a snapshot of the stock dataset with all four sentiment scores. Table 16 shows all of the findings from stock market prediction using linear SVC derived for the four sentiment scores. Here, accuracy, mean absolute error (MAE), and mean square error (MSE) are initially assessed for stock data alone without any sentiment score. Each of these four sentiment scores is then combined incrementally, and the results vary. Gradually,

Table 12 First category experimental results in terms of accuracy, recall, and F1 score

	Stock data with ADS-1			Stock data with ADS-2			Stock data with ADS-3			Stock data with ADS-4		
	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 score
VADER	95.68%	82.14%	90.16%	94.82%	78.57%	88%	96.55%	89.28%	92.59%	97.23%	90.78%	94.12%
Logistic Regression	97.44%	91.28%	94.78%	93.48%	77.72%	87.65%	96.97%	89.75%	90.19%	95.96%	83.33%	91.26%
Loughran–McDonald	95.08%	80.65%	89.17%	95.59%	81.45%	89.76%	95.99%	82.81%	90.77%	96.35%	85.71%	92.3%
Henry	94.57%	78.20%	87.68%	93.39%	77.34%	87.29%	97.08%	90.48%	93.87%	96.23%	85.62%	92.21%
TextBlob	96.09%	85.35%	91.72%	96.13%	85.52%	92.11%	95.92%	83.27%	91.41%	96.02%	85.11%	91.34%
Linear SVC	97.29%	90.98%	94.43%	96.80%	90.12%	93.02%	96.73%	89.77%	92.98%	98.11%	91.62%	95.18%
Stanford	94.13%	77.68%	86.89%	96.72%	89.62%	92.88%	94.3%	77.51%	87.12%	93.75%	78.52%	88.61%

accuracy improves while error drops. Here, in Table 16, there is one exception when stock data are combined with ADS-1_SVC. A spike is shown in this case. Otherwise, all other cases followed the incremental pattern.

Additionally, Table 15 illustrates that public opinion and news articles/headlines have an effect on stock market movement forecast performance. The cumulative influence of these characteristics improves prediction accuracy while lowering the cost. The following figures illustrate the prediction graphs and the accompanying cost versus epoch graph. The cost function and prediction graph in Fig. 7a and b are shown without any emotion score. The subsequent figures (Figs. 8, 9, 10 and 11 with (a) and (b) counterparts) demonstrate that when sentiment ratings are integrated sequentially, accuracy increases and cost decreases. Figure 10b illustrates the optimal prediction when all four sentiment ratings are added together.

7 Comparison with existing works

Table 17 shows the comparison of this proposed research work with two of the latest published related works. As indicated in the table, the current work scrapes a sufficient number of online sources (four sources) to perform sentiment analysis using seven tools. In this work, prediction operations are carried out using the LSTM deep learning model. When four sentiment scores and stock data are combined, the current proposed work achieves a high degree of accuracy. Thus, when four sentiment scores from seven sentiment tools are merged with stock data, the experimental setup described in Table 2 results in a higher prediction accuracy.

Dutta et al. 2021 used the Vader sentiment analysis tool on news articles from the "Economic Times", and an LSTM deep learning model was implemented on the BSE stock index. Since the work is related to the proposed model, which addresses the NSE stock index and six more sentiment analysis tools to perform sentiment analysis on four different data sources, we compared it to represent the enhancement in the performance by enhancing the number of features.

Wang et al. 2021 considered six stock prices (extra features included in this paper are adjusted close and volume) with sentiment scores of news headlines. Here, the vaderSentiment library is used for sentiment analysis as one of our sentiment analysis tools, and a total of six machine learning approaches, SVM, neural networks, naïve Bayes-based method, and random forest, logistic regression and XGBoost model, were tested. In future work, they mentioned using a deep learning approach. The authors also indicated the limited amount of news articles collected. As the proposed work is related and

Table 13 Second category experimental results in terms of cost functions

Tool	Metric	Stock data with ADS-1 and ADS-2	Stock data with ADS-1, ADS-2 and ADS-3	Stock data with ADS-1, ADS-2, ADS-3 and ADS-4
Henry	R-squared	0.2361	0.1285	0.2749
	MSE	0.0215	0.0391	0.0252
	MAE	0.0894	0.0966	0.0877
Logistic regression	R-squared	0.2159	0.1834	0.3798
	MSE	0.0242	0.031	0.0123
	MAE	0.0907	0.0925	0.0809
Loughran–McDonald	R-squared	0.2107	0.2091	0.2118
	MSE	0.0253	0.0272	0.0243
	MAE	0.0901	0.0911	0.0899
VADER	R-squared	0.1767	0.2363	0.3375
	MSE	0.0379	0.0204	0.0193
	MAE	0.0968	0.0897	0.0813
TextBlob	R-squared	0.2982	0.2818	0.2472
	MSE	0.0215	0.0243	0.0259
	MAE	0.0847	0.09	0.0825
Linear SVC	R-squared	0.2233	0.2603	0.435
	MSE	0.0329	0.0271	0.0103
	MAE	0.0902	0.089	0.0789
Stanford	R-squared	0.1881	0.2461	0.2739
	MSE	0.0322	0.0272	0.0259
	MAE	0.0943	0.0869	0.0852

Table 14 Second category experimental results in terms of accuracy, recall, and F1 score

Accuracy, recall and F1-score for combination data	Stock data with ADS-1 and ADS-2			Stock data with ADS-1, ADS-2 and ADS-3			Stock data with ADS-1, ADS-2, ADS-3 and ADS-4		
Tools for calculating Sentiment scores	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score
VADER	95.87%	83.36%	91.78%	94.87%	79.5%	87.84%	96.85%	88.76%	91.69%
Logistic Regression	94.67%	81.57%	88.51%	94.25%	78.87%	87.45%	97.67%	91.13%	93.72%
Loughran–McDonald	94.75%	81.81%	88.97%	94.63%	79.27%	87.48%	94.78%	81.98%	89.02%
Henry	94.8%	82.11%	89.62%	92.99%	76.88%	84.69%	96.36%	87.53%	90.94%
TextBlob	96.76%	88.42%	90.83%	96.28%	87.05%	90.21%	96.48%	88.15%	91%
Linear SVC	96.15%	86.85%	89.82%	96.35%	87.41%	90.73%	98.32%	93.12%	95.24%
Stanford	93.62%	78.74%	86.82%	96.46%	87.8%	91.11%	96.57%	87.95%	91.4%

we tried to overcome the limitation as well as incorporate the future work of this paper here, we compared the performance with the performance of the proposed work to indicate improvement.

Table 18 has also been included to provide all four data combination results with stock data from each sentiment analysis tool (seven sentiment analysis tools) for both classification and regression. We found that in all the metrics, linear SCV provides the best performance.

8 Conclusion and future work

The purpose of this research is to predict and analyse stock market movement during a lockdown situation caused by the COVID-19 outbreak using sentiment scores. Four internet sources are used to scrape data in this case, including "Stock Market Related News Headlines", "Twitter's Tweets", "Financial News Articles" and "Facebook Comments". Seven sentiment analysis tools are utilized to determine the sentiment scores of four web scraped datasets: logistic

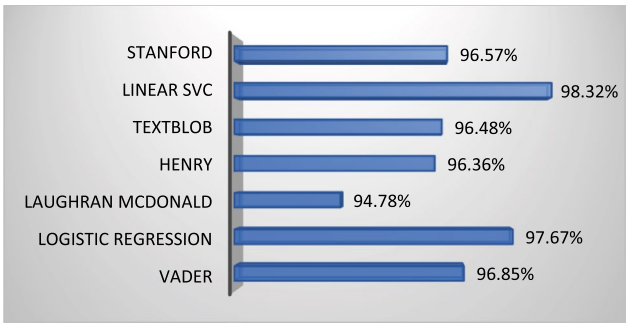


Fig. 6 Comparative performance from each tool with combined datasets

regression, linear support vector classifier, Vader, Stanford's Core-NLP, Textblob, Henry, and Loughran–McDonald. Each of the seven tools on the four sources provides 28 sets of sentiment scores on an average daily basis. These scores are paired with stock data in two categories to conduct two types of Stock Market Movement Prediction tests in regression and classification.

The first category of the experiment associates stock data with individual sentiment scores and forecasts the stock market using an LSTM deep learning network. The accuracy of a prediction is expressed as a percentage. In the second category of the experiment, sentiment scores from four online sources for each tool are gradually integrated with stock data to assess the combined influence of all four sentiment scores on prediction performance. The following conclusions may be drawn from the outcomes of these tests:

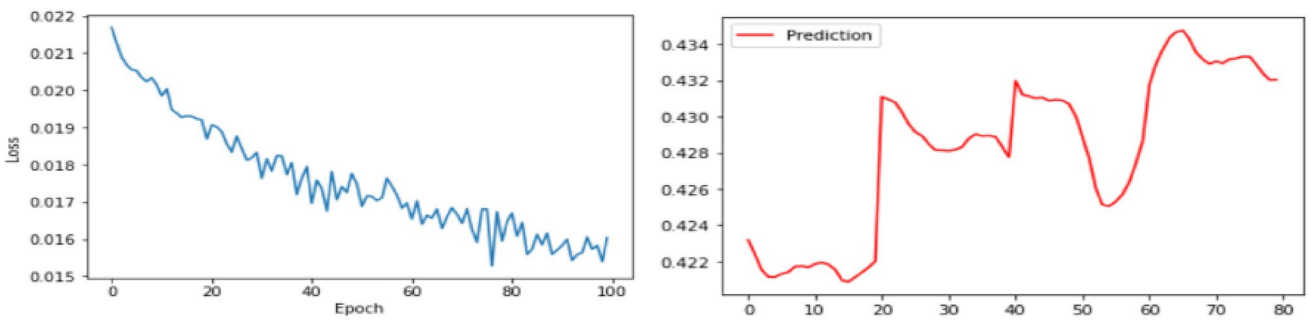
The highest percentage accuracy is reached when the average sentiment ratings of Facebook comments derived using linear SVC are paired with stock data. Here, the accuracy is 98.11 percent.

Table 15 Sample Nifty50 stock data with all four linear SVC sentiment scores

Date	Open	High	Low	Close	ADS-1_SVC	ADS-2_SVC	ADS-3_SVC	ADS-4_SVC	Class
07–01-2020	10,323.79	10,447.04	10,299.59	10,430.04	0.75	0.192	1	1	1
07–02-2020	10,493.04	10,598.20	10,485.54	10,551.70	1	0.049	0.068	0.068	1
07–03-2020	10,614.95	10,631.29	10,562.65	10,607.34	0	0.069	0.078	0.078	0
07–06-2020	10,723.84	10,811.40	10,695.09	10,763.65	0.667	0.139	0.139	0.139	1
07–07-2020	10,802.84	10,813.79	10,689.70	10,799.65	0.25	0.106	-0.667	0.154	0

Table 16 Stock market prediction results with linear SVC's four sentiment scores

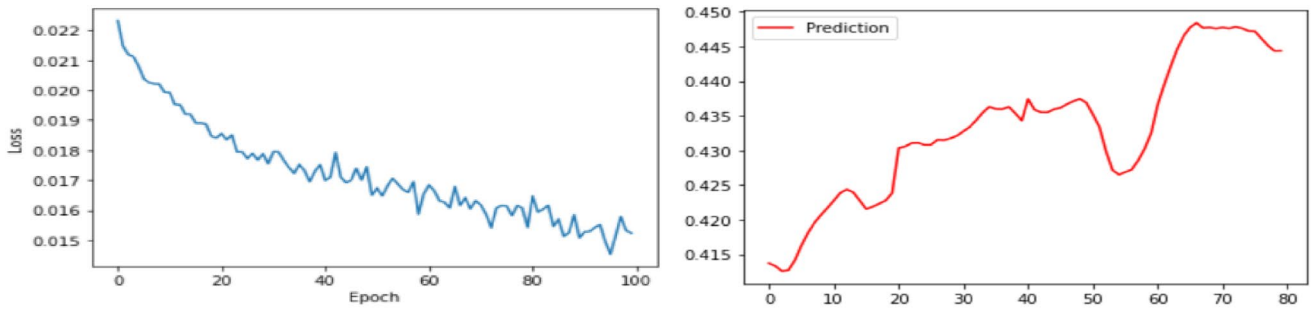
Results	MAE	MSE	Accuracy
Dataset			
Stock Data Only	0.097	0.0545	95.22%
Stock Data + ADS-1_SVC	0.0809	0.0115	97.29%
Stock Data + ADS-1_SVC + ADS-2_SVC	0.0902	0.0329	96.15%
Stock Data + ADS-1_SVC + ADS-2_SVC + ADS-3_SVC	0.089	0.0271	96.35%
Stock Data + ADS-1_SVC + ADS-2_SVC + ADS-3_SVC + ADS-4_SVC	0.0789	0.0103	98.32%



(a) Loss vs epoch graph

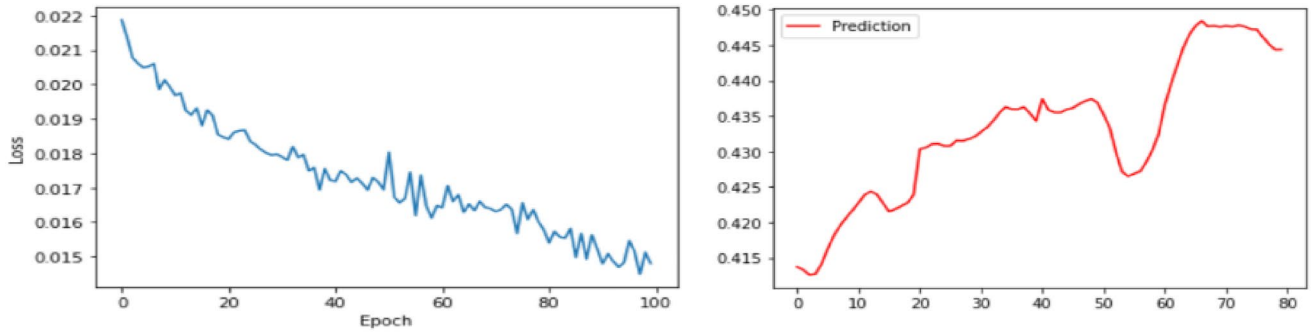
(b) Prediction graph: Prediction vs days

Fig. 7 Stock market movement prediction without any sentiment score



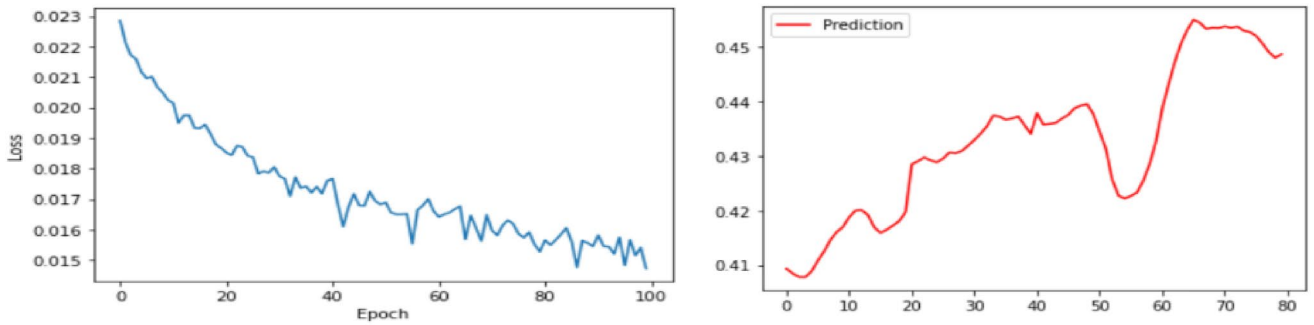
(a) Loss vs epoch graph (b) Prediction graph: Prediction vs days

Fig. 8 Stock market movement prediction with one sentiment score (news headlines) from linear SVC



(a) Loss vs epoch graph (b) Prediction graph: Prediction vs days

Fig. 9 Stock market movement prediction with two sentiment scores (news headlines and Twitter) from linear SVC

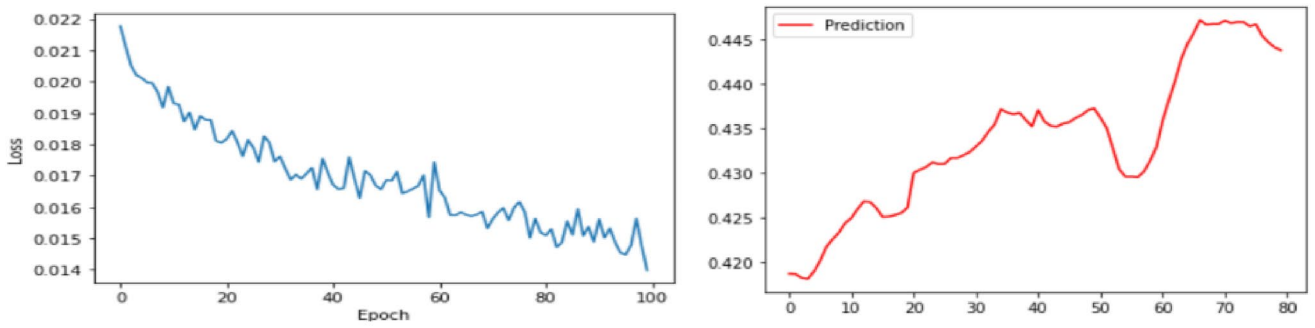


(a) Loss vs epoch graph (b) Prediction graph: Prediction vs days

Fig. 10 Stock market movement prediction with three sentiment scores (news headlines, Twitter and news articles) from linear SVC

When average sentiment scores from four sources are incrementally integrated with stock data from each tool, five tools' scores give the greatest performance when combined with all four sentiment scores. The best performance is achieved by the composite impact of linear SVC-generated sentiment ratings, which is 98.32 percent.

The overall analysis of the results reveals that using public mood and news headlines/articles in combination with stock data improves forecast accuracy. As a result, increasing the number of sentiments used during the lockdown period improves forecast accuracy.



(a) Loss vs epoch graph (b) Prediction graph: Prediction vs days

Fig. 11 Stock Market Movement Prediction with four sentiment scores (news headlines, Twitter, news articles, and Facebook comments) from Linear SVC

Table 17 Comparison with two existing works

Works done	Online data sources for sentiment analysis	Sentiment analysis tools	Stock price prediction method	Stock market data	Accuracy	
(Dutta, Pooja, Jain, Panda, and Nagwani, 2021)	News articles from some newspapers like "Economic Times"	VADER	LSTM	S & P 500 from Yahoo Finance	77.45%	
(Wang et al. 2021)	Stock related news headlines from online media sources like "The New York Times"	VADER	Machine Learning Algorithms	Dow Jones Industrial Average (DJIA) from Yahoo Finance	72.98%	
Proposed work	Stock related articles headlines from "Economic Times," Tweets from Twitter, Financial news from "Economic Times" and Facebook comments	VADER, Logistic Regression, Loughran–McDonald, Henry, TextBlob, Linear SVC and Stanford	LSTM	Nifty50 (NSE) from Yahoo Finance	Linear SVC	98.32%
					Logistic Regression	97.67%
					VADER	96.85%
					Loughran–McDonald	94.78%
					Henry	96.36%
					TextBlob	96.48%
					Stanford	96.57%

Table 18 Comparison with two existing works

Experimental results for combination data	Stock data with ADS-1, ADS-2, ADS-3 and ADS-4					
Experiment categories	Classification result			Regression result		
Tools for calculating sentiment scores	Accuracy	Recall	F1 score	R-squared	MSE	MAE
VADER	96.85%	88.76%	91.69%	0.3375	0.0193	0.0813
Logistic regression	97.67%	91.13%	93.72%	0.3798	0.0123	0.0809
Loughran–McDonald	94.78%	81.98%	89.02%	0.2118	0.0243	0.0899
Henry	96.36%	87.53%	90.94%	0.2749	0.0252	0.0877
TextBlob	96.48%	88.15%	91%	0.2472	0.0259	0.0825
Linear SVC	98.32%	93.12%	95.24%	0.435	0.0103	0.0789
Stanford	96.57%	87.95%	91.4%	0.2739	0.0259	0.0852

Linear SVC-generated sentiment scores from "Facebook comments" produce the best performances.

This study may be extended in the future to include additional stock indexes and the use of deep learning as another sentiment analysis tool to monitor changes in forecast accuracy. Stock market technical indicators are another parameter that can be used in conjunction with stock data to assess the accuracy of stock market movement forecasts.

Author contributions ND and SC provided conceptualization; ND and TC did methodology; ND and BS done formal analysis and investigation; ND performed writing—original draft preparation; BS was involved in writing—review and editing; SC contributed to supervision.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Code availability All codes for data cleaning and analysis associated with the current submission are available.

References

Akter, S., Aziz, M. T. (2016) Sentiment analysis on facebook group using lexicon based approach. In: 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT), 1–4. doi: https://doi.org/10.1109/CEEICT.2016.7873080

Arif MH, Li J, Iqbal M, Liu K (2018) Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Comput 22(21):7281–7291. https://doi.org/10.1007/s00500-017-2729-x

Batra, R., Daudpota, S. M. (2018) Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In: 2018 international conference on computing, mathematics and engineering technologies (ICoMET), 1–5. Sukkur: IEEE. doi: https://doi.org/10.1109/ICOMET.2018.8346382

Biswas S, Ghosh A, Chakraborty S, Roy S, Bose R (2020) Scope of sentiment analysis on news articles regarding stock market and GDP in struggling economic condition. Int J Emerg Trends Eng Res 8(7):3594–3609. https://doi.org/10.30534/ijeter/2020/117872020

Bonta V, Kumaresh N, Janardhan N (2019) A comprehensive study on lexicon based approaches for sentiment analysis. Asian J Comput Sci Technol 8(S2):1–6

Budiharto W (2021) Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). J Big Data 8(1):47. https://doi.org/10.1186/s40537-021-00430-0

Chaudhuri, A., Mukherjee, S., Chowdhury, S., Sadhukhan, B., Goswami, R. T. (2018). Fractality and Stationarity Analysis on Stock Market. In: 2018 international conference on advances in computing, communication control and networking (ICACCCN), 395–398. Greater Noida (UP), India: IEEE. doi: https://doi.org/10.1109/ICACCCN.2018.8748504

Chauhan P, Sharma N, Sikka G (2021) The emergence of social media data and sentiment analysis in election prediction. J Ambient Intell Humaniz Comput 12(2):2601–2627. https://doi.org/10.1007/s12652-020-02423-y

Chou, C., Park, J., Chou, E. (2021) Predicting Stock Closing Price After COVID-19 Based on Sentiment Analysis and LSTM. In: 2021 IEEE 5th advanced information technology, electronic and automation control conference (IAEAC), 5, 2752–2756. doi: https://doi.org/10.1109/IAEAC50856.2021.9390845

Derakhshan A, Beigy H (2019) Sentiment analysis on stock social media for stock price movement prediction. Eng Appl Artif Intell 85:569–578. https://doi.org/10.1016/j.engappai.2019.07.002

Dutta A, Pooja G, Jain N, Panda RR, Nagwani NK (2021) A hybrid deep learning approach for stock price prediction. In: Joshi A, Khosravy M, Gupta N (eds) Machine learning for predictive analysis. Springer, Singapore, pp 1–10

Eck M, Germani J, Sharma N, Seitz J, Ramdasi PP (2021) Prediction of stock market performance based on financial news articles and their classification. In: Sharma N, Chakrabarti A, Balas VE, Martinovic J (eds) Data management, analytics and innovation. Springer, Singapore, pp 35–44

Elena, P. (2021). Predicting the movement direction of omxs30 stock index using xgboost and sentiment analysis. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:bth-21119

Gers FA, Schmidhuber E (2001) LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Trans Neural Netw 12(6):1333–1340. https://doi.org/10.1109/72.963769

Ghasiya P, Okamura K (2021) Understanding the Middle East through the eyes of Japan's Newspapers: a topic modelling and sentiment analysis approach. Digit Scholarsh Humanit 36(4):871–885. https://doi.org/10.1093/llc/fqab019

Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N (2021) Explainable stock prices prediction from financial news articles using sentiment analysis. Peer J Comput Sci 7:e340. https://doi.org/10.7717/peerj-cs.340

Hajhmida MB, Oueslati O (2021) Predicting mobile application breakout using sentiment analysis of Facebook posts. J Inf Sci 47(4):502–516. https://doi.org/10.1177/0165551520917099

Hassan S-U, Aljohani NR, Idrees N, Sarwar R, Nawaz R, Martínez-Cámara E, Herrera F (2020) Predicting literature's early impact with sentiment analysis in Twitter. Knowl-Based Syst 192:105383. https://doi.org/10.1016/j.knosys.2019.105383

Hassan, O. A.-H., Ramaswamy, L., Miller, J. A. (2009). MACE: A dynamic caching framework for mashups. In: *2009* IEEE international conference on web services, 75–82. Los Angeles, CA, USA: IEEE. https://doi.org/10.1109/ICWS.2009.119

Huang, Z., Tanaka, F. (2021). MSPM: A modularized and scalable multi-agent reinforcement learning-based system for financial portfolio management. http://arxiv.org/abs/2102.03502 [Cs, q-Fin]. Retrieved from

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Zhang, J. (2021) LSTM Based sentiment analysis for cryptocurrency prediction. http://arxiv.org/abs/2103.14804 [Cs]. Retrieved from

Hussein DME-DM (2018) A survey on sentiment analysis challenges. J King Saud Univ - Eng Sci 30(4):330–338. https://doi.org/10.1016/j.jksues.2016.04.002

Kordonis, J., Symeonidis, S., Arampatzis, A. (2016) Stock price forecasting via sentiment analysis on twitter. In: Proceedings of the 20th pan-hellenic conference on informatics, 1–6. New York, NY,

USA: Association for Computing Machinery https://doi.org/10.1145/3003733.3003787
Lin, B., Zampetti, F., Bavota, G., Di Penta, M., Lanza, M., Oliveto, R. (2018) Sentiment analysis for software engineering: How far can we go? In: Proceedings of the 40th international conference on software engineering, 94–104. Gothenburg Sweden: ACM https://doi.org/10.1145/3180155.3180195
Lu Y, Zheng Q (2021) Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic. Curr Issue Tour 24(7):892–898. https://doi.org/10.1080/13683500.2020.1843607
Ly, T. H., Nguyen, K. (2020). Do words matter: predicting ipo performance from prospectus sentiment. In: 2020 IEEE 14th international conference on semantic computing (ICSC), 307–310. https://doi.org/10.1109/ICSC.2020.00061
Marengo D, Azucar D, Longobardi C, Settanni M (2021) Mining facebook data for quality of life assessment. Behav Inform Technol 40(6):597–607. https://doi.org/10.1080/0144929X.2019.1711454
Mehta P, Pandya S, Kotecha K (2021) Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. Peer J Comput Sci 7:e476. https://doi.org/10.7717/peerj-cs.476
Mokhtari S, Yen KK, Liu J (2021) Effectiveness of artificial intelligence in stock market prediction based on machine learning. Int J Comput Appl 183(7):1–8. https://doi.org/10.5120/ijca2021921347
Mukherjee S, Sadhukhan B, Sarkar N, Roy D, De S (2021) Stock market prediction using deep learning algorithms. CAAI Trans Intell Technol. https://doi.org/10.1049/cit2.12059
Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB (2020) Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. J Am Acad Dermatol 83(3):803–808. https://doi.org/10.1016/j.jaad.2019.07.014
Patel JM (2020) Getting structured data from the internet: running web crawlers/scrapers on a big data production scale. Springer, Berkeley
Rajput A (2020) Chapter 3—natural language processing, sentiment analysis, and clinical analytics. In: Lytras MD, Sarirete A (eds) Innovation in health informatics. Academic Press, Cambridge, pp 79–97
Rase MO (2020) Sentiment analysis of Afaan Oromoo facebook media using deep learning approach. New Med Mass Commun. https://doi.org/10.7176/NMMC/90-02
Shang Y, Li H, Zhang R (2021) Effects of pandemic outbreak on economies: evidence from business history context. Front Public Health 9:146. https://doi.org/10.3389/fpubh.2021.632043
Shi Y, Zheng Y, Guo K, Ren X (2021) Stock movement prediction with sentiment analysis based on deep learning networks. Concurr Comput: Pract Exp 33(6):e6076. https://doi.org/10.1002/cpe.6076
Singh M, Jakhar AK, Pandey S (2021) Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc Netw Anal Min 11(1):33. https://doi.org/10.1007/s13278-021-00737-z
De S Sirisuriya, S. C. M. (2015). *A* Comparative study on web scraping. Retrieved from http://ir.kdu.ac.lk/handle/345/1051
Turner Z, Labille K, Gauch S (2021) Lexicon-based sentiment analysis for stock movement prediction. J Constr Mater. https://doi.org/10.36756/JCM.v2.3.5
Valle-Cruz D, Fernandez-Cortez V, López-Chau A, Sandoval-Almazán R (2021) Does twitter affect stock market decisions? Financial sentiment analysis during pandemics: a comparative study of the H1N1 and the COVID-19 periods. Cogn Comput. https://doi.org/10.1007/s12559-021-09819-8
Van Houdt G, Mosquera C, Nápoles G (2020) A review on the long short-term memory model. Artif Intell Rev 53(8):5929–5955. https://doi.org/10.1007/s10462-020-09838-1
Wagner AF (2020) What the stock market tells us about the post-COVID-19 world. Nat Hum Behav 4(5):440–440. https://doi.org/10.1038/s41562-020-0869-y
Wang, Z., Hu, Z., Li, F., Ho, S.-B. (2021) Learning-based stock market trending analysis by incorporating social media sentiment analysis [Preprint]. In Review. doi: https://doi.org/10.21203/rs.3.rs-181424/v1
Wu S, Liu Y, Zou Z, Weng T-H (2021) S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. Connect Sci. https://doi.org/10.1080/09540091.2021.1940101