Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Studies of Sentiment Analysis for Stock Market Prediction using Machine Learning: A Survey towards New Research Direction

Joyeta Chandra[1*], Abhoy Chand Mondal[2]

[1]Research Scholar, Department of Computer Science, The University of Burdwan, India
[2]Professor, Department of Computer Science, The University of Burdwan, India

DOI: https://doi.org/10.36347/sjet.2025.v13i01.007 | **Received:** 07.12.2024 | **Accepted:** 16.01.2025 | **Published:** 21.01.2025

***Corresponding author:** Joyeta Chandra
Research Scholar, Department of Computer Science, The University of Burdwan, India

Abstract

Original Research Article

In an era defined by economic ups and downs, as well as the general loss in purchasing power parity, many are utilizing additional passive income streams to protect from inflation. Due to the peak of digitization and an increase in online services than offline, investing in the stock market is now preferred by people as a passive income. But investing in the stock market has risks that these can result in great financial losses. One of the risks posed by this volatility is that driven by investor sentiment, a single piece of news can go viral completely rightly - or misleadingly wrong, and its impact on an individual's investment decisions would be all too significant to influence in which direction trends may shift within financial markets. This paper addresses the evolving field of sentiment analysis for understanding stock market dynamics using machine learning methods. It provides a comprehensive review of the current state-of-the-art literature in terms of research approaches, datasets and results with respect to works published from 2018 till early 2023. It describes a varity of machine learning models, sentiment lexicons and data sources needed to analyze sentiments and how they affect us when it comes to predictions in stock markets. Furthermore, this study serves to explain the workings of sentiment and its role in market movements or so it appears when a narrative is told about stock trends from such perspective that we compile here.

Keywords: Sentiment analysis, Machine learning techniques, Stock market analysis, Investor sentiment, Behavioral finance, Literature review.

1. INTRODUCTION

In recent years, academic focus has shifted towards investors' sentiment and its impact on market perfor-mance. This sentiment mirrors future expectations of cash flows and risk [23]. Traditional theories ignored its role, but it drives stock price fluctuations and uncertainty. Changes in India's financial landscape have increased participant diversity and influenced risk-taking behaviors. Over the past two decades, Internet technologies reshaped India's stock market. Online access removed barriers, allowing investors to trade shares anywhere, anytime [14]. In an inflationary environment, many view the stock market as a passive income source, despite the associated risks. This has led to losses for inexperienced traders within a short span.

Sentiment, defined as the expression of positive, neutral, or negative emotions in text or spoken language, plays a crucial role in today's digital era. Its significance is especially pronounced in the context of sentiment analysis, a technique used to decipher emotions within text, which is pivotal in understanding public opinion [12]. This method has gained prominence with the growth of social media, making sentiment analysis a critical tool for predicting stock market movements. Sentiment analysis has been transformed due to Machine Learning. Multiple early ML algorithms, such as Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, allowed systems to learn sentiment pattens from labelled datasets [27]. However, with the increasing volume, velocity, and variety of data-commonly known as three Vs of big data – demand for scalable, high-performance solutions has become a priority. Deep Learning (DL) and transformer-based architectures, in particuler, have advanced the capabilities of SA in big data constellations. The (LSTM) networks and Convolutional Neural Networks (CNN) are deep learning architectures that extract the detailed, contextual, sequential data datasets. Transformer Model (e.g. BERT, GPT) inherent in massive data is transforms big data into state of the art NLP systems, using big a dimension of processing distributed compute [28, 29].

Renowned for its tumultuous, dynamic, and nonlinear character, the stock market defies predictability. Research demonstrates that stock prices aren't completely random and can be predicted to some extent [16]. Understanding stock patterns and external factors is crucial for enhancing price forecasts. Various elements, including investor reactions to financial news, daily events and rumors, contribute to stock market price fluctuations. Various social media platforms like YouTube, Twitter, and Facebook, often propagate rumors or false news about companies to manipulate their audience. Stocktwits, a platform designed for those interested in the stock market, also plays a role. Users share market insights, stock ideas, and real-time information there. The market operates under the principle of "buy on rumors, sell on news" [5], reflecting how sentiment can influence trading decisions.

The intent of this paper is to (i) comprehensively review existing literature, (ii) elaborate on the motivations, methodologies, datasets, and limitations identified within these studies, and (iii) pinpoint directions for future research. The following is how this document is structured: An analytical evaluation of numerous works in the field of sentiment analysis (SA) applied to machine learning (ML)-based stock market predictions are given in Section 2. Section 3 discusses the papers reviewed, presenting the findings through an appropriatetabular format. Finally, Section 4 offers conclusions and outlines potential avenues for future investigation.

2. METHODOLOGY

A. Research Question (RQ)

1) Which Stock Market oriented dataset and Sentiment analysis dataset they have worked on? This question aimed to find the stock market and sentiment corpus collecting platform names on which maximum and minimum work has already been done.
2) What kind of approaches they applied for their work and what results they achieved? This second question aims to survey which models or approaches they have applied to carry out their desired work and tries to points out the results that have been achieved.
3) Identify the sub - application area of the reviewed paper. Third question tried to categories the reviewed papers by identifying the sub-application areas.

B. Search and Selection Parameters

Various online journal databases, such as Scopus, IEEE, Science Direct, ACM Digital Library, Web of Science, SpringerLink, and Google Scholar are utilized as sources to conduct this review. The initial search consisted the following string: ("Sentiment Analysis" OR "Sentimental Analysis" OR "Opinion Mining") AND ("Stock market" OR "Stock market data" OR "Share market" OR ”Stock Index") AND (”Machine learning" OR ”Machine learning technique") with the publication year ranges from 2018 to 2023. After gathering an adequate amount of literature, the documents underwent a refining process, prioritizing what to incorporate and what to omit (Table 1).

Table 1: Inclusion and Exclusion criteria

Selection Criteria	Journal database
Inclusion criteria of documents	Peer-reviewed journals, research papers and articles, like reputed conference papers accepted manuscripts.
Exclusion criteria of documents	Non-English documents, documents including missing abstracts during the study and generic and irrelevant documents, documents that are not peer-reviewed, review papers.

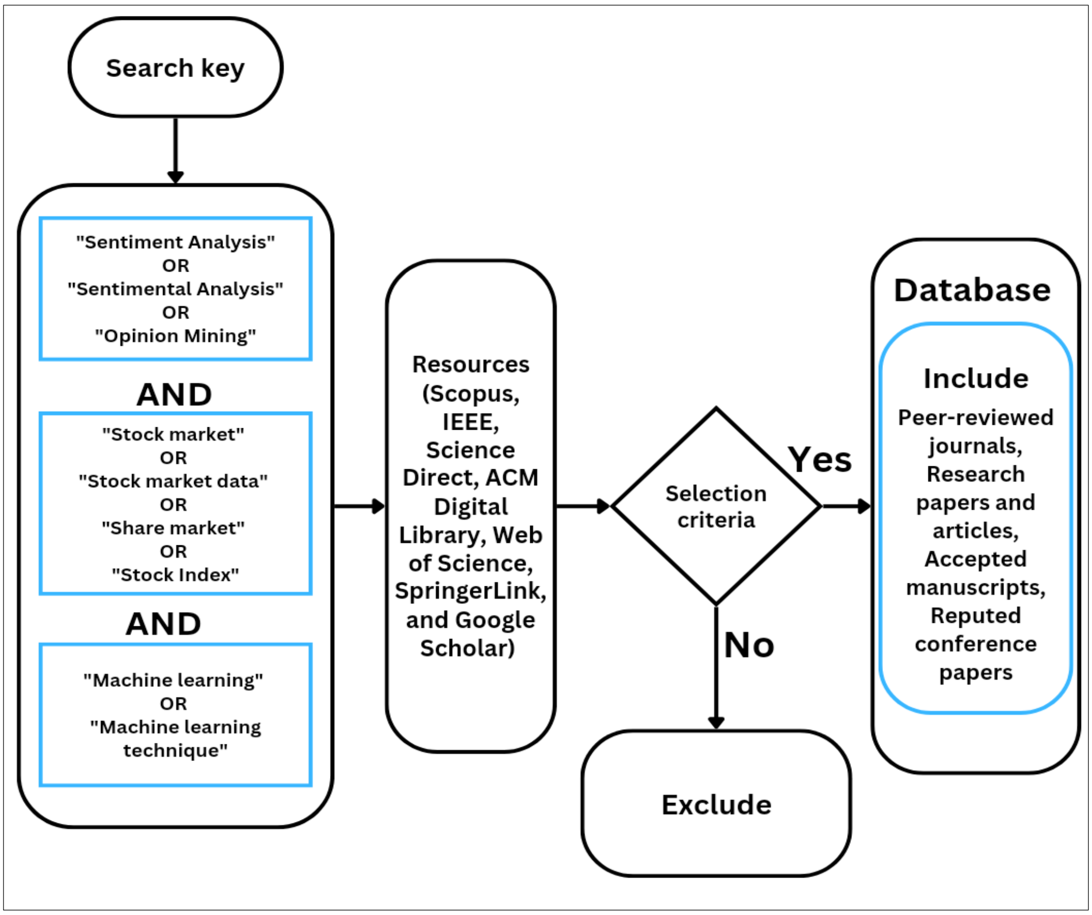
Search key
"Sentiment Analysis"
OR
"Sentimental Analysis"
OR
"Opinion Mining"
AND
"Stock market"
OR
"Stock market data"
OR
"Share market"
OR
"Stock Index"
AND
"Machine learning"
OR
"Machine learning technique"
Resources (Scopus, IEEE, Science Direct, ACM Digital Library, Web of Science, SpringerLink, and Google Scholar)
Selection criteria
Yes
No
Exclude
Database
Include
Peer-reviewed journals, Research papers and articles, Accepted manuscripts, Reputed conference papers

Figure 1: Literature selection process flowchart

3. LITERATURE REVIEW

3.1 Stock Price Prediction

Previous SA studies on stock price prediction linked movements with news articles, but faced limitations due to insufficient textual data, affecting predictive accuracy. To address this, Mohan *et al*., [13], gathered extensive time series data alongside relevant news articles. They examined 265,463 news articles and the daily closing prices of S&P 500 companies between February 2013 and March 2017. Various forecasting models, including ARIMA, RNN, and Facebook Prophet, were rigorously evaluated. Despite challenges with low or volatile stock prices, RNN models exhibited promise in correlating textual content with stock price direction.

In pursuit of forecasting stock market movement, Boukatif *et al*., [3], propose an ML methodology encompassing five key steps. They have gathered historical stock price data for ten NASDAQ-100 companies and advocate for a more precise approach through sentiment and texture features. Employing N-grams for finer-grained textual analysis, we aim to enhance predictive capability. Various learning algorithms, including Logistic Regression (LR), Na¨ıve Bayes (NB), SVM, XGBoost, Random Forest (RF), and ANN,were applied to training data. Their model demonstrates promising accuracy, achieving a 60% prediction rate.

The MS-SSA-LSTM (Multi-source - Sparrow Search Algorithm - Long short term memory) model was presented by Mu *et al*., [15] for the purpose of predicting stock prices. To train and evaluate the model, they chose six sample datasets from the Chinese financial sector. The MS-SSA-LSTM model combines deep learning methods, swarm intelligence algorithms, and SA with multi-source data that affects stock prices. Comparative analysis against standard LSTM reveals a notable improvement, with an average R^2 enhancement of 10.74%.

The efficiency of an attention-based LSTM deep neural network in forecasting future stock market movements is examined by Xu *et al*., [26]. They create both individual and aggregate datasets using technical indicators for 80 US stocks, stock history data, and SA of financial tweets (a sizable dataset obtained from StockTwits that spans the period from January 1, 2016, to December 31, 2018). Comparing conventional LSTM with attention-based LSTM, the latter demonstrates improved performance on the aggregate dataset. Particularly intriguing are the results from the individual stock dataset, with the best achieving approximately 65% accuracy, indicating promising potential.

Batra *et al*., [2], evaluate the StockTwits integration with SA for enhancement of stock price prediction accuracy. By analyzing tweets related to Apple products from StockTwits and corresponding market data from Yahoo Finance (2010-2017), they created a comprehensive dataset. They attain note worthy training (75.22%) and test (76.68%) accuracies by employing SVM (Support Vector Machine), which was selected for its accuracy and resilience in text categorization.

Gupta *et al*., [7], examined SA on StockTwits tweets for 5 major companies: General Electric, Apple, Microsoft, Amazon, and Target. StockTwits, a microblogging platform, served as a key data source, with a specialized downloader developed for extensive tweet collection. Yahoo Finance provided stock price data. Nine months of analysis took place between January 1st and September 30th, 2019, with 120 days dedicated to training and the remaining time to testing. Employing three ML methods (NB, SVM, and LR) as well as 5 featurization methodologies, LR with TF-IDF achieved notable accuracy (75%-85%), emphasizing the efficacy of SA in stock price prediction.

Mehta *et al*., [10], investigated if ARIMA, LSTM, as well as Linear Regression work for predicting stock pricesin a short-term perspective. The Nifty 50 index has been taken up to analyze the performance of three major shares included within it Bharti Airtel, State Bank of India, and Bosch Limited. To bring public sentiment into our analysis, we collected tweets regarding individual stocks using Tweepy, a Python library for Twitter data collection. The results showed that ARIMA is better than LSTM for short-term predictions and DNA demonstrated higher accuracy in long-term forecasting than its rivals. After that, because our use case calls for daily model execution and we have a short-range forecast error analys is approach, the lower RMSE score of ARIMA allowed us to conclude it was more adequate according to their specific situation.

A SA and prediction system was developed by Gondaliya *et al*., [6], for the Indian stock market around the COVID-19 epidemic. Utilizing data gathered from Twitter and news websites between January 1st, 2020, and August 24th, 2020, they compared six ML algorithms utilizing both Bag-of-Words and TF-IDF approaches. Their findings show that Bag-of-Words outperformed TF-IDF in terms of sentiment classification accuracy in five of the six algorithms: Decision Tree, LR, NB, RF, and SVM. The top performances with Bag-of-Words were SVM and LR, both with 78% accuracy.

TAJMAZINANI *et al*., [24], performed research investigating the influence of technical and fundamental analysis on Iranian stocks. By utilizing the HESNEGAR lexicon, the researchers analyzed the sentiment of news articles related to these stocks. A deep learning model, CNN (Convolution Neural Network), was applied for the prediction of stock performance depending on three approaches that are price only, news sentiment only, and acombination of both. The results revealed that the hybrid model, incorporating both technical indicators as well as news sentiment, achieved superior prediction accuracy compared to the additional approaches, representing the efficiency of a comprehensive approach for Iranian stock market trading strategies.

In order to forecast stock values, Padmanayana *et al*., [17], combined SA of news headlines and Twitter tweets with historical stock data. XGBoost was used as the ML model, with data obtained from FinViz and Twitter using scraping and Tweepy, respectively. Vader Sentiment Analyzer was employed for SA. When it came to forecasting the stock prices of firms for example Amazon, Apple, and Microsoft, the model's accuracy was 89.8%.

3.2 Stock Market Trend Prediction

Minh *et al*., [11], suggested the TGRU (Two-stream Gated Recurrent Unit), achieving an accuracy of 66.32%, surpassing GRU and LSTM models. TGRU’s bidirectional learning states enhance data absorption, especially in text processing. Using financial data and the Harvard IV-4 sentiment dictionary, they created a sentiment Stock2Vec embedding, which they demonstrated outperformed Glove and Word2Vec. Two tests were carried out: one used historical data and articles from Reuters/Bloomberg to predict the stock price directions of the S&P 500 index, while the other used Viet Stock news and cophieu68 stock prices to forecast the trends of the VN-index. Data spanned from October 9th to November 9th, 2017, showcasing TGRU’s efficiency in the prediction of stock price.

Qiu *et al*., [20], introduce a unique weighting method for stock reviews, using data from Estimony. com, a financial website, focusing on the SSE 50 (Shanghai Stock Exchange 50). We include the day-of-the-week and holiday influences to improve the sentiment index's authenticity and dependability. Finally, the effectiveness of this index, compared to a daily sentiment index, is evaluated utilizing eight ML models (DT, SVM, GBDT, RF, LR, AdaBoost, NB, and KNN). The results show significant improvement in prediction accuracy across all models with the modified sentiment index, with KNN and SVM experiencing the most notable gains (12.25% and 68.37%, respectively).

In the work of Koukaras *et al*., [9], aims for the enhancement of stock prediction accuracy by integrating various SA and ML methodologies. They collected tweets from Twitter and Stock-Twits, along with financial data from Yahoo Finance for Microsoft stock. The tweets underwent SA, and seven ML models were tested: SVM, KNN(K-Nearest Neighbors), NB, LR, Decision Tree (DT), MLP (MultilayerPerceptron), and

RF. The Valence Aware Dictionary and sEntiment Reasoner (VADER) for SA in conjunction with SVM produced the greatest results, with an F-score of 76.3% and an Area Under Curve (AUC) value of 67%.

A hybrid Recurrent Neural Network (HyRNN) architecture that incorporates Bidirectional Long Short-Term Memory (Bi-LSTM), GRU, and sLSTM is proposed by John *et al*., [8], to enhance stock price predictions. By integrating financial news sentiments from NASDAQ INLUDING stock features, the HyRNN outperforms the RNN-GRU model. The hybrid model achieved a lower MAE (14.62 vs. 15.7 for stock features alone, and 13.1 vs. 14.3 with sentiment data). The coefficient of determination (R^2) was also higher for HyRNN (0.987 vs. 0.983).

The study of Ren *et al*., [21], suggested a refined sentiment index that addresses the day-of-week effect. By adjusting sentiment indexes based on past weekend changes and generalizing to holidays by presenting anexponential function that weights recent sentiment changes on weekends more heavily than older changes, they improve their accuracy. They collected financial review data from Sina Finance and Eastmoney. Using SVM, an ML model, we predicted the SSE 50 Index, a significant Chinese stock market index. Their empirical investigation shows that when it comes to predicting market direction, combining sentiment traits with stock market data performs noticeably better than utilizing stock market data alone. The accuracy increased by 18.6%, reaching 89.93%, after incorporating sentiment variables.

The study by Carosia *et al*., [4], of seeks to predict movements in the Brazilian financial market by analyzing Twitter sentiment during the 2018 presidential elections (1^{st}Octto 31^{st}Dec2018). The methodology includes the development of an SA module for Portuguese, followed by a comparison of ML methodologies (NB, Maximum Entropy, SVM, and MLP) to identify more suitable model for the financial domain. Unlabeled tweets gathered between October 1st, 2018, and December 12th, 2018 were subjected to the SA module. Three methods were used for the analysis: sentiment weighted by Retweets (RTs), sentiment weighted by Favorites (FAVs), and daily sentiment count. MLP was the most successful approach for SA in Portuguese, according to the study's findings, which also examined the relationship between financial market movement and the emotion that predominated on social media.

3.3 Stock Market Volatility prediction

Archary *et al*., [1], gathered technical features spanning a decade to feed into an RNN model, employing different optimization methodologies for predicting the stock volatility of Apple (AAPL), Amazon (AMZN), and Microsoft (MSFT). They conducted a time series analysis from August 1, 2009, to August 1, 2019, encompassing 2516 trading days. Integrating both content and technical historical datasets, the RNN identified correlations between data points, successfully predicting trend directions, despite seemingly below-average results. Their study underscores the efficacy of RNNs in stock volatility prediction, emphasizing the importance of combining diverse datasets for improved forecasting accuracy.

3.4 Cryptocurrency Price Prediction

Valencia *et al*., [25], suggest a computational approach for prediction price movement on the transaction pairs of Litecoin, Ripple, Ethereum, and Bitcoin using common ML tools along with social media data. We acquired social data from Twitter (an average of 345,000 tweets each day; a total of 20789572 tweets) and historical market data via the Cryptocompare.com public api. For input, they used three models: Neural Networks (NN), SVM, and RF aiming to compare the prediction quality of them. The first one means social data alone, the second represents market data only, and the third combines both. Neural Networks performed the best among all three when using both Twitter and market data as inputs.

Another research conducted by Pant *et al*., [18], proposed for the prediction of the volatile Bitcoin price by analyzing Twitter sentiment and looking for the connection between these two. From various news accounts, such as CryptoCurrency (@cryptocurrency), BitcoinNews (@BTCTN), BitcoinMagazine (@BitcoinMagazine), CryptoYoda (@CryptoYoda1338), Bitcoin Forum (@BitcoinForums), CoinDesk (@coindesk), and Roger Ver(@rogerkver), they collected Bitcoin-related tweets that were posted from January 1, 2015 to December 31,2017. Historical price data for the same period was sourced from Coin market cap. For sentiment classification, they developed a voting classifier comprising NB, Linear SVM, and RF, achieving an accuracy of 81.39% with a validation split of 1:3. The classified tweets and historical price data were then fed into an RNN model, which predicted the next time frame’s price with an accuracy of 77.62%.

Serafini *et al*.'s paper [22], suggests a deep learning and statistical model to forecast the daily weighted price of Bitcoin. After training ARIMAX and LSTM-based RNN models with different combinations of emotion and financial characteristics, we discovered that the best combination just includes the weighted price of Bitcoin and the sentiment expressed in tweets. Their analysis demonstrates that the linear ARIMAX model, despite its simplicity, outperforms the complex LSTM-based RNN regarding prediction accuracy, achieving a mean squared error of 0.00030187 on new predictions.

4. RESEARCH QUESTION ANALYSIS

RQ1: Which Stock Market Oriented Dataset and Sentiment Analysis Dataset They Have Worked on?

According to the reading of the research papers, there are basically two type of dataset. First, they choose the index or companies that they want to work on and gather the financial data of that selected index or companies. Second, there are data about that previously chosen index or companies from various social media sites or news articles which is used for sentiment analysis. For the financial dataset the papers [1-17], choose the indexes or companies that belongs to US stock market. The papers [19, 10], [6], choose Indian stock market index and paper [15-21], [20], work on Chinese financial market. Except these three there are some other [24, 4], paper which use other country stock market data. Beside the traditional stock market dataset [25, 22], carried out their work on virtual currency, popular as Cryptocurrency. The following pie chart (Figure 2) shows the distribution of stock market dataset of the reviewed literature.

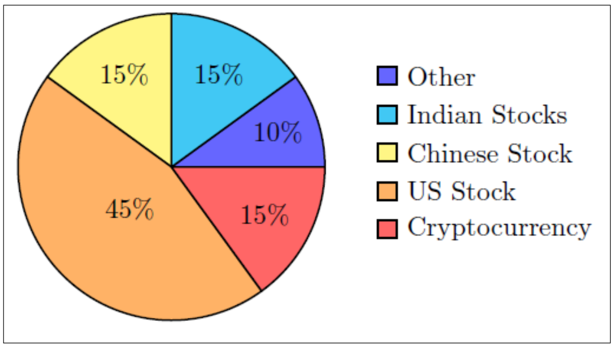

Figure 2: Distribution of Stock market dataset

For collecting dataset for sentiment classification most of the papers [1-25], choose Twitter. Except Twitter, the papers [2-20], use other platform like StockTwits, various other Microbloging website. Some papers [6-22], also use news articles for the purpose either beside Twitter or alone. The distribution of sentiment classification dataset of the reviewed literature is represented in the form of pie chart (Figure 3) bellow.

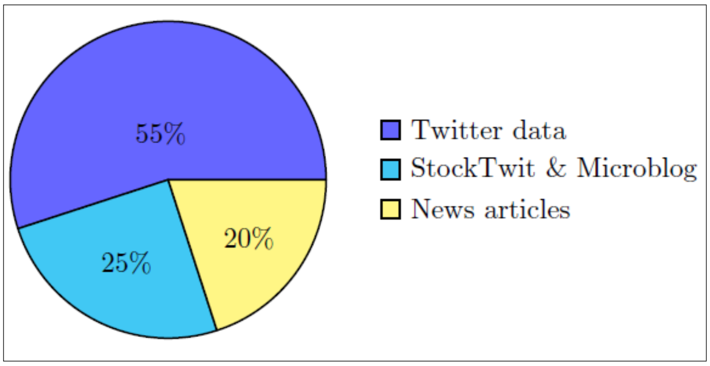

Figure 3: Distribution of Sentiment analysis dataset

RQ2: What Kind of Approaches They Applied for Their Work and What Results They Achieved?

As the answer of this question, we have organized the key findings of our literature review into the Table 2 bellow. This table summarizes the overall approach carried out and the results achieved of each study.

Table 2: Summary of Applied Approaches and the results of all Literature reviewed papers

Reference	Applied approachs	Result achieved
Archary *et al*., (2020) [1]	Integrate both content and technical historical datasets and feed to the RNN which identified correlations between data points	Successfully done the prediction
Batra *et al*., (2018) [2]	Employ SVM for both sentiment classification and stock movement prediction	Achieved accuracy of 75.22% in training and 76.68% in testing.
Bukatif *et al*., (2020)[3]	Na¨ıve Bayes, LR, SVM, ANN, Random forest, and XGBoost were applied associated with N-grams for finer-grained textual analysis	Models achieved 60% accuracy

Reference	**Applied approachs**	**Result achieved**
Carosia *et al.*, (2020)[4]	The analysis was carried out using three approaches: sentiment weighted by Retweets (RTs), sentiment weighted by Favorites (FAVs), and daily sentiment count and after that four ML techniques (Naive Bayes, Maximum Entropy, SVM, and Multilayer Perceptron) for comparison.	Multilayer Perceptron outperform others.
Gondaliy *et al.*, (2021)[6]	Compared six machine learning algorithms utilizing both Bag-of-Words and TF-IDF approaches.	SVM and LR achieved the highest accuracy of 78% using Bag-of-Words features.
Gupta *et al.*, (2020)[7]	Employing three ML methods- NB, SVM, LR and five featurization techniques	Logistic regression with TF-IDF achieved notable accuracy of 75%-85%
John *et al.*, (2023)[8]	Propose a hybrid Recurrent Neural Network (HyRNN) architecture that combining Bi-LSTM, GRU and sLSTM	HyRNN outperforms the RNN-GRU model.
Koukaras *et al.*, (2022)[9]	Tested seven ML models - K-Nearest Neighbors (KNN), Na¨ıve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Multilayer Perceptron (MLP).	VADER with SVM achieved an F-score of 76.3% and AUC of 67%.
Mehta *et al.*, (2021)[10]	Investigated if ARIMA, LSTM and Linear Regression work for predicting stock prices in a short-term perspective	ARIMA outperform other two models acquiring lower RMSE score.
Minh *et al.*, (2018)[11]	Proposed a Two-stream Gated Recurrent Unit with bidirectional learning states and developed Sentiment Stock2Vec that incorporates financial data and sentiment analysis using the Harvard IV-4 sentiment dictionary.	Achieved an accuracy of 66.32%, outperforming traditional GRU and LSTM models.
Mohan *et al.*, (2019)[13]	Incorporat both historical stock price data and textual information from financial news articles. The models employed include traditional time series models like ARIMA and Facebook Prophet, as well as more complex recurrent neural networks (RNNs) with LSTM architectures.	RNN models, especially those incorporating textual information, Outperformed traditional time series models.
Mu *et al.*, (2023)[15]	Developed MS-SSA-LSTM (Multi-source – Sparrow Search Algorithm - Long short term memory) model for stock price prediction.	R^2 of MS-SSA-LSTM is improved by 10.74% on average compared to standard LSTM.
Padmanayana *et al.*, (2021)[17]	Employed Vader Sentiment Analyzer followed by XGBoost for stock price prediction.	Achived accuracy of 89.8%.
Pant *et al.*, (2018)[18]	A voting classifier, comprising Naive Bayes, SVM, Random Forest, developed for sentiment classification (validation split: 1:3). Classified tweets and historical prices were used as input for an RNN model.	Voting classifier achieved 81.39% accuracy for sentiment classification and RNN model, which predicted the next time frame's price achieved an accuracy of 77.62%.
Reference	**Applied approachs**	**Result achieved**
Qiu *et al.*, (2022) [20]	weighting method incorporating with the day-of-the-week and holiday effects and then apply SVM, DT, RF, GBDT, NB, LR, AdaBoost, and KNN to evaluate the predictive performance of the modified sentiment index.	Average accuracy of all models improved by 7.28%, in particular KNN and SVM increases of 12.25% and 68.37%, respectively.
Ren *et al.*, (2018)[21]	Adjusted sentiment indexes of holidays using exponential function and then employed SVM for stock trend prediction.	Achieved accuracy of 89.93% with a rise of 18.6%.
Serafini *et al.*, (2020) [22]	Trained ARIMAX and LSTM-based RNN models with various combinations of financial and sentiment features.	ARIMAX model outperforms LSTM-based RNN in terms of prediction accuracy with mean squared error of 0.00030187.
Tajmazinani *et al.*, (2022) [24]	Utilizes HESNEGAR lexicon for sentiment classification afterward CNN was employed to predict stock performance using three distinct approaches: a price-only model, a news sentimentonly model, and a hybrid model that incorporated both price and news sentiment data.	The hybrid model, incorporating both technical and fundamental factors, demonstrated superior predictive capabilities.
Valencia *et al.*, (2019)[25]	Neural Networks (NN), Support Vector Machines (SVM) and Random Forests (RF) aiming to compare the prediction quality of them. NN model used solely social data as input, while the SVM model focused exclusively on market data. The RF model combined both social and market data to assess their combined predictive power.	Neural Networks performed the best among all three when using both Twitter and market data as inputs.
Xu *et al.*, (2019)[26]	Construct both individual and aggregate datasets comprising stock history data, financial tweets sentiment and technical indicators and applied on attention-based LSTM.	Attention-based LSTM acheived 65% accuracy as the best.

RQ3: Identify the Application Area of the Reviewed Paper

After reviewing the 20 papers, we've identified four major sub-application fields. The Table 3 presents the references and the number of papers in each field.

Table 3: Identified Sub-application areas and the papers for each categories

Application area	Reference	No. of papers
Stock Market Volatility prediction	[1]	1
Cryptocurrency Price prediction	[18] [22] [25]	3
Stock Market Trend prediction	[4] [8] [9] [11] [20] [21]	6
Stock Market Price prediction	[2] [3] [6] [7] [15] [10] [13] [17] [24] [26]	10

5. DISCUSSION

The reviewed studies collectively highlight the growing trend of leveraging social media sentiment analysis and Machine learning and Deep learning techniques for stock market prediction. All the three questions mentioned and analyzed in the above section are designed to understand the insights of each papers especially their dataset, their approaches and their application areas.

RQ 1 shows the supremacy of US stock market dataset (45%) when it comes to work in stock market. Except US stock market there are others stock market oriented works are also done but the percentage of them are quite low. That is basically because of the easy availability of data in US stocks than other stock market datasets like Chinese stock market dataset (15%), Indian stock market dataset (15%) and others (10%). Besides the traditional stock market datasets prediction of Cryptocurrency price are also comes in light (15%). On the other hand most research favored platforms like Twitter (55%) and StockTwits or various other Microblogging websites (25%) for capturing sentiment analysis data, reflecting their significance in capturing real-time market sentiment. Besides these, analysing News articles (20%) are also a very traditional way for creating sentiment analysis dataset.

According to the analysis of RQ 2, the reviewed papers use various ML or various DL models or hybrid models. After incorporating sentiment analysis component, performance all type of models have improved. But on an average the highest accuracy that is achieved is 85% - 89% so far.

As the analysis of RQ 3 indicates, we have identified four sub application areas under the broad umbrella of Stock market prediction. Among the four categories, most of the work has done in Stock market price prediction sub-application area and stock market volatility prediction has quite less works. Stock market trend prediction sub-application area comes in second place. Except these three, Cryptocurrency price prediction becomes another sub-application area as it is becoming popular to trade in virtual currency.

6. CONCLUSION & FUTURE DIRECTION

This paper presents a comprehensive exploration of how sentiment analysis, leveraged through various machine learning techniques, can significantly influence stock market predictions. By examining diverse datasets, from social media platforms like Twitter and StockTwits to traditional news outlets, the studies reviewed underline the pivotal role of investor sentiment in forecasting stock movements. Despite the challenges posed by the vast and dynamic nature of data, the application of models ranging from Naive Bayes, SVM and RNN, CNN to advanced and complex Hybrid models demonstrates promising results in deciphering the complex relationship between public sentiment and market trends. This analysis of existing work has also unveiled three primary research gaps. First, there is a notable deficiency in research targeting the other stock market rather than US stock market. Second, the peak accuracy achieved by current predictive models is 85% - 89% so far. Third stock market volatility prediction field has the potential for more work.

Addressing these gaps, future research should pivot towards enhancing the accuracy of predictive models by incorporating more sophisticated techniques and considering the unique aspects of different stock markets, such as India's. Moreover, the increasing participation of younger investors highlights the need for models that can dynamically adapt to shifts in market demographics and sentiment sources. By doing so, the financial community can develop more nuanced and effective tools for navigating the complexities of global stock markets, ultimately leading to more informed investment decisions and potentially higher returns.

Our survey, focused on machine learning for stock market sentiment analysis, reviews 20 papers. We acknowledge the possibility of missing relevant studies, accepting this as a limitation of our work.

REFERENCES

1. Archary, D., & Coetzee, M. (2020). Predicting stock price movement with social media and deep learning. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–5

2. Batra, R., & Daudpota, S. M. (2018). Integrating Stocktwits with sentiment analysis for better prediction of stock price movement. *2018 International Conference on Computing, Mathematics and Engineering Technologies (ICoMET),* IEEE, 1–5
3. Bouktif, S., Fiaz, A., & Awad, M. (2020). Augmented textual features-based stock market prediction. *IEEE Access, 8*, 40269–40282.
4. Carosia, A. E. D. O., Coelho, G. P., & Silva, A. E. A. D. (2020). Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. *Applied Artificial Intelligence, 34*(1), 1–19.
5. Deveikyte, J., Geman, H., Piccari, C., & Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence, 5,* Article 836809.
6. Gondaliya, C., Patel, A., & Shah, T. (2021). Sentiment analysis and prediction of Indian stock market amid COVID-19 pandemic. *IOP Conference Series: Materials Science and Engineering, 1020,* Article 012023.
7. Gupta, R., & Chen, M. (2020). Sentiment analysis for stock price prediction. *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),* IEEE, 213–218.
8. John, A., & Latha, T. (2023). Stock market prediction based on deep hybrid RNN model and sentiment analysis. *Automatika, 64*(4), 981–995.
9. Koukaras, P., Nousi, C., & Tjortjis, C. (2022). Stock market prediction using microblogging sentiment analysis and machine learning. *Telecom, 3*, 358–378.
10. Mehta, Y., Malhar, A., & Shankarmani, R. (2021). Stock price prediction using machine learning and sentiment analysis. *2021 2nd International Conference for Emerging Technology (INCET)*, IEEE, 1–4.
11. Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access, 6*, 55392–55404.
12. Mohammad, S. M. (2021). Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion Measurement* (pp. 323–379). Elsevier.
13. Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService),* IEEE, 205–208.
14. Mohnot, R., Mohnot, R., & Muralidharan, P. (2017). Role of electronic trading in emerging stock markets performance. *International Journal of Business Forecasting and Marketing Intelligence, 3*(4), 388–406.
15. Mu, G., Gao, N., Wang, Y., & Dai, L. (2023). A stock price prediction model based on investor sentiment and optimized deep learning. *IEEE Access*.
16. Mudinas, A., Zhang, D., & Levene, M. (2019). Market trend prediction using sentiment analysis: Lessons learned and paths forward. *arXiv preprint arXiv:1903.05440*. https://arxiv.org/abs/1903.05440
17. Padmanayana, V., & Bhavya, K. (2021). Stock market prediction using Twitter sentiment analysis. *International Journal of Scientific Research in Computer Science, Engineering, and Information Technology.*
18. Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018). Recurrent neural network-based Bitcoin price prediction by Twitter sentiment analysis. *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, IEEE, 128–132.
19. Pathak, A., & Shetty, N. P. (2019). Indian stock market prediction using machine learning and sentiment analysis. In *Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM 2017* (pp. 595–603). Springer.
20. Qiu, Y., Song, Z., & Chen, Z. (2022). Short-term stock trends prediction based on sentiment analysis and machine learning. *Soft Computing, 26*(5), 2209–2224.
21. Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal, 13*(1), 760–770.
22. Serafini, G., Yi, P., Zhang, Q., Brambilla, M., Wang, J., Hu, Y., & Li, B. (2020). Sentiment-driven price prediction of Bitcoin based on statistical and deep learning approaches. *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 1–8.
23. Shen, J., Yu, J., & Zhao, S. (2017). Investor sentiment and economic forces. *Journal of Monetary Economics, 86*, 1–21.
24. Tajmazinani, M., Hassani, H., Raei, R., & Rouhani, S. (2022). Modeling stock price movements prediction based on news sentiment analysis and deep learning. *Annals of Financial Economics, 17*(1), Article 2250003.
25. Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy, 21*(6), Article 589.
26. Xu, Y., & Keselj, V. (2019). Stock prediction using deep learning and sentiment analysis. *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 5573–5580.
27. Banerjee, S., & Mondal, A. C. (2024). An intelligent approach towards plant leaf disease detection through different convolutional neural networks.

International Journal of Intelligent Systems and Applications in Engineering, 12(2), 536–546.

28. Banerjee, S., & Mondal, A. C. (2024). A sophisticated approach to soil productivity detection using a convolutional neural network-based model. *International Journal of Advanced and Applied Sciences, 11*(8), 198–210.
29. Banerjee, S., & Mandal, A. C. (2024). An innovative approach involves machine learning algorithms to forecast future farmer revenue. *SSRG International Journal of Electrical and Electronics Engineering, 11*(8), 59–71. https://doi.org/10.14445/23488379/IJEEE-V11I8P106
30. Banerjee, S., Das, S., & Mondal, A. C. (2024). Classification of healthy and diseased broccoli leaves using a custom deep learning CNN model. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12*(5), 110–116. https://doi.org/10.55524/ijircst.2024.12.5.15
31. Banerjee, S., Palsani, D., & Mondal, A. C. (2024). Nutritional content detection using vision transformers: An intelligent approach. *International Journal of Innovative Research in Engineering and Management (IJIREM), 11*(6), 21–27. https://doi.org/10.55524/ijirem.2024.11.6.3
32. Banerjee, S., Das, S., & Mondal, A. C. (2024). A study of the application domain of large language models in the agricultural sector. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST), 12*(5), 74–78. https://doi.org/10.55524/ijircst.2024.12.5.10
33. Banerjee, S., & Mondal, A. C. (2023). An ingenious method for estimating future crop prices that emphasises machine learning and deep learning models. *International Journal of Information Technology (Singapore), 15*, 4291–4313. https://doi.org/10.1007/s41870-023-01474-6