Neural Computing and Applications (2025) 37:10771–10794
https://doi.org/10.1007/s00521-024-10807-w

S.I.: TIMELY ADVANCES OF DEEP LEARNING WITH APPLICATIONS AND DATA DRIVEN MODELING

Forecasting stock market volatility using social media sentiment analysis

Christina Saravanos[1] · Andreas Kanavos[2]

Received: 30 October 2023 / Accepted: 16 November 2024 / Published online: 13 December 2024

Abstract
In the era where social media significantly influences public sentiment, platforms such as Twitter have become vital in predicting stock market trends. This paper presents a cutting-edge predictive model that integrates historical stock market data, Twitter sentiment analysis, and an extensive array of tweet-related features. Utilizing advanced regression techniques and deep neural networks, our model forecasts the daily closing prices of the U.S. stock market indices with notable accuracy and demonstrates a strong link between market values, sentiment scores, and social media activities. Our analysis particularly emphasizes the importance of tweet diffusion and the influence of prominent Twitter users in refining prediction accuracy. Contrary to conventional wisdom, we discovered that incorporating a wide range of tweet-derived features significantly improves the model's performance without leading to sparsity challenges. This study not only questions established paradigms but also underscores the potential of social media analytics in financial market forecasting, with substantial implications for investors, market analysts, and policy makers.

Keywords Stock market forecasting · Sentiment analysis · Natural language processing (NLP) · Social media analytics · Regression techniques · Deep neural networks (DNNs)

1 Introduction

Social media platforms have transformed into pivotal arenas for communication, interaction, and the dissemination of information to vast audiences. Among these platforms, Twitter distinguishes itself through its unique medium of *tweets*-concise messages that rapidly disseminate across user networks, often encapsulating the emotions or opinions of their authors [36]. This swift propagation of sentiment through tweets significantly influences public perception and viewpoint, impacting various sectors, including the stock market [76].

The challenge of accurately predicting stock market volatility is exacerbated by the dynamic and often noisy flow of information. Market fluctuations are influenced by a myriad of external factors, including geopolitical events, financial crises, and shifts in public perception. In response, recent years have seen the ascendancy of regression techniques and deep neural networks (DNNs) in forecasting market movements. These methodologies leverage historical data and sentiment analysis from social media to detect complex patterns, promising advances in prediction accuracy despite challenges posed by overfitting and noise inherent in traditional regression methods and the high dimensionality faced by DNNs [16].

This paper proposes an innovative approach to estimating the daily closing prices of the U.S. stock market indices by harmonizing regression techniques with deep neural networks. By integrating historical stock prices, sentiment derived from social media, and a variety of

✉ Andreas Kanavos
akanavos@ionio.gr

Christina Saravanos
saravanou@ceid.upatras.gr

[1] Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece

[2] Department of Informatics, Ionian University, 49100 Corfu, Greece

tweet-based features, our method not only establishes a strong correlation between actual and predicted daily close prices but also highlights the enhancement in predictive accuracy afforded by considering metadata and the social presence of tweet authors. Crucially, our analysis reveals the profound impact of tweet volume and diffusion in predicting stock market trends, challenging conventional wisdom regarding the influence of sparse features on model performance.

The contributions of this paper are manifold:

- Integration of multiple data sources: We introduce a comprehensive analytic framework that synthesizes historical stock data, Twitter sentiment analysis, and an array of tweet-derived features to forecast daily closing prices of the U.S. stock indices.
- Synergistic application of regression techniques and DNNs: Our research employs a synergistic blend of regression models and deep neural networks, showcasing the enhanced predictive capabilities of our approach.
- Elucidating the impact of social media sentiment: The study highlights the pivotal role of social media sentiment, particularly the influence of tweet diffusion and the digital footprint of authors, in accurately forecasting stock market dynamics.
- Redefining analytical paradigms: We challenge entrenched beliefs about the detrimental effects of feature sparsity on predictive models, offering evidence to the contrary.

These contributions advance our understanding of the interplay between social media analysis and stock market forecasting. By merging historical stock data with the sentiment-rich environment of social media, particularly Twitter, our study unveils the potential of public sentiment as a key indicator of market trends. This approach not only broadens the scope of predictive analytics in finance but also underscores the value of combining diverse data sources, including both structured historical financial data and unstructured social media content. Our findings challenge prevailing assumptions about model performance in the face of feature sparsity, offering new insights into effective predictive model construction. This work paves the way for innovative forecasting methodologies that leverage the vast data generated in our digital society, marking a significant step forward in financial analysis and prediction.

The remainder of the paper is organized as follows: Sect. 2 reviews pertinent literature on stock market prediction techniques. Section 3 details the proposed analytic framework, including data sources, feature extraction, and the application of regression techniques and DNNs. Section 4 outlines the experimental setup, followed by Sect. 5 which discusses the results. The paper concludes with Sect. 6, summarizing our findings and their implications.

2 Literature review

The stock market's sensitivity to a range of external factors, including public sentiment, political events, and trends in social media and search engines, complicates the prediction of market volatility. The inherently volatile, dynamic, and noisy data from the stock market add further complexity to forecasting efforts [65, 66].

Recent research has introduced several innovative paradigms for predicting stock market volatility or price movements. Traditional regression techniques such as linear regression [4, 11], auto-regression [18], support vector regression (SVR) [38, 40], random forest [21, 59, 67], extra tree [59], and XGBoost regression [57, 82] have been widely applied. Concurrently, deep neural networks (DNNs) including artificial neural networks (ANNs), recurrent neural networks (RNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs) have emerged as effective tools for managing the large volumes and complex nature of stock market data [4]. Notably, RNN variants such as long short-term memory (LSTM) networks and gated recurrent units (GRUs) have been specifically employed to forecast U.S. stock market volatility [61], while CNNs have been used to predict future prices in the Chinese and Indian stock markets [15, 50]. GANs have been proposed for price prediction in indices like the FTSE MIB, CSI 300, and S&P 500 [72, 73, 86]. Hybrid models that combine neural networks with traditional machine learning techniques have also shown promise in enhancing prediction accuracy [45, 56, 83].

Social media platforms, due to their wide user base and diverse communication services, play a crucial role in estimating stock market volatility [37]. Innovations in this area have leveraged the correlation between historical stock prices and sentiment expressed in related social media posts. Groundbreaking work has analyzed the relationship between stock market volatility and social media sentiment, particularly focusing on Indian and U.S. stock markets [26, 55]. These studies have significantly advanced our understanding of how social media sentiment impacts stock market trends. Moreover, the application of auto-regression models like the auto-regressive moving average (ARIMA) model in conjunction with social media analysis has opened new avenues for predicting stock market movements.

Emerging techniques for forecasting U.S. stock market volatility have emphasized the utility of correlating public sentiment, as derived from online posts, with historical

stock prices, utilizing advanced LSTMs or CNNs for analysis. Sentiment analysis is achieved through a variety of techniques, including natural language processing (NLP), classification algorithms, and word embeddings [8, 35]. Hybrid models combining neural networks, such as LSTMs and CNNs, have proved effective in capturing the volatility of markets by analyzing sentiment from online posts, demonstrating considerable promise in accurately predicting future market trends through the integration of deep learning and NLP.

This review underscores the diverse and evolving landscape of stock market prediction research, highlighting the transition toward more integrative and sophisticated analytical models that leverage both traditional financial data and the rich sentiment information available through social media.

3 Overview of the presented scheme

This section details the innovative scheme developed to forecast the daily closing prices of a U.S. stock market index. As illustrated in Fig. 1, our approach is structured around three pivotal modules, each designed to harness distinct sets of data and analytical techniques, culminating in a comprehensive predictive model.

3.1 Module 1: Feature Extraction and Sentiment Analysis from Twitter

The initial phase of our methodology involves a detailed analysis of Twitter data. This module is tasked with extracting a multifaceted set of features from tweets, which include metadata attributes, the social media engagement metrics of authors, and textual analysis leveraging natural language processing (NLP) techniques. Key NLP operations, such as n-grams and Parts-of-Speech (POS) tagging, are applied to the tweet text to gauge sentiment and objectivity, providing a nuanced understanding of public sentiment as it relates to stock market movements.

3.2 Module 2: Stock Market Data Feature Extraction

Parallel to the analysis of Twitter data, this module focuses on extracting crucial features from historical stock market data pertaining to the target index, sourced from Yahoo! Finance. The extraction process is facilitated by the Yahoo! Finance API, ensuring the retrieval of accurate and timely financial data critical for prediction accuracy.

3.3 Module 3: Prediction of Daily Close Prices

Integrating the insights gained from the preceding modules, the final module employs a sophisticated blend of regression techniques and deep neural networks to predict the daily closing prices of the target stock market index. This

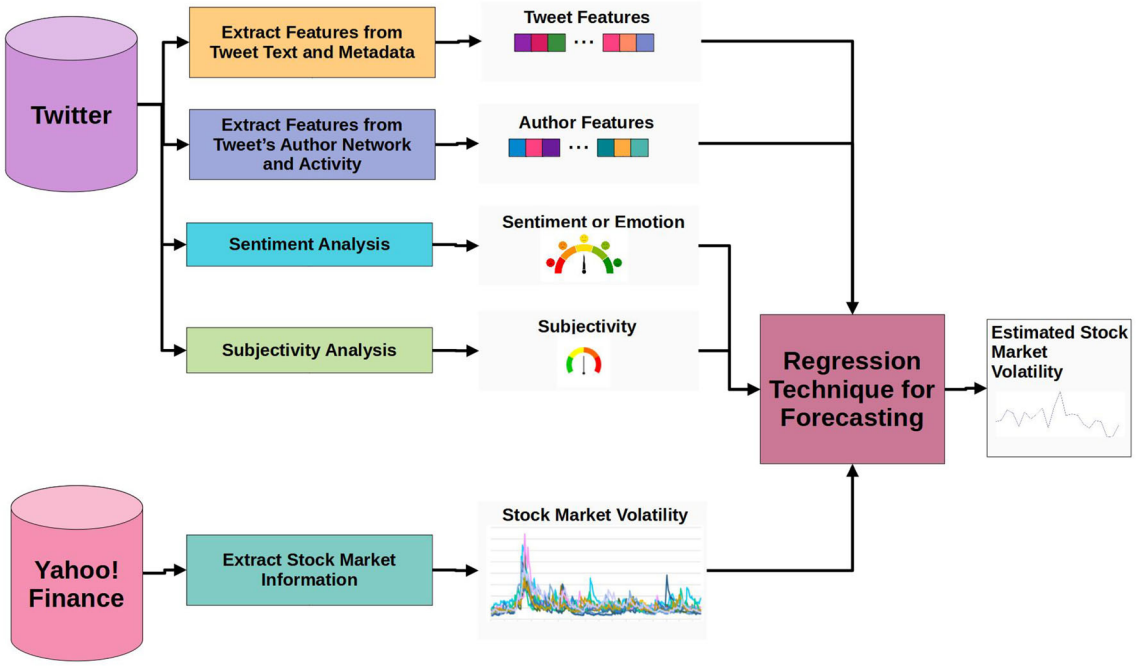

Fig. 1 Schematic representation of the predictive scheme

comprehensive approach enables the model to identify and leverage significant patterns within the aggregated datasets, leading to the generation of precise and reliable market predictions.

The described scheme represents a cutting-edge integration of social media sentiment analysis and traditional financial data analysis, utilizing state-of-the-art machine learning techniques. This holistic approach not only enhances the accuracy of stock market predictions but also offers insights into the complex dynamics that influence market trends, standing as a testament to the potential of combining diverse data sources and advanced analytical methodologies in financial forecasting.

3.4 Features

This subsection delineates the comprehensive array of features extracted from the preliminary two modules, which are instrumental in the estimation and prediction of daily closing prices of a U.S. stock market index. These features are meticulously curated from varied sources, encompassing stock market data from Yahoo! Finance, metadata and textual analysis of tweets, as well as an evaluation of the tweet authors' social media presence.

3.4.1 Stock Market Features

We incorporate a wide spectrum of historical data points, including price trends and trading volumes. These features are vital for understanding market behavior and trends over time, serving as indicators of potential market movements.

3.4.2 Twitter Metadata Features

Metadata extracted from tweets offers a wealth of information about user interactions and behaviors. This includes publication times of tweets, user engagement metrics such as likes, retweets, and replies, as well as user-specific information like follower counts. Such metadata provides insights into the relevance and impact of tweets, reflecting public interest and sentiment at given times.

3.4.3 Textual Content Features

The application of natural language processing (NLP) techniques to tweet content enables the extraction of sentiment scores, subjectivity levels, and semantic patterns. Techniques such as sentiment analysis, POS tagging, and named entity recognition are employed to discern the underlying sentiment and viewpoints regarding the stock market index. This analysis offers a nuanced perspective on public sentiment, augmenting the prediction model with the emotional and opinionated undertones of market-related discussions [34].

3.4.4 Integration and Analysis

These multifaceted features are integrated within our predictive model, undergoing rigorous analysis to discern patterns and correlations with the stock market index's daily closing prices. The synthesis of traditional financial indicators with nuanced social media analytics facilitates a comprehensive understanding of the factors driving market fluctuations. Through this approach, our model leverages both quantitative financial data and qualitative sentiment analysis, enabling a robust and informed forecasting of market trends.

The strategic selection and incorporation of these features underscore our model's capacity to capture the complex dynamics affecting the stock market, thus enhancing the accuracy and reliability of our predictions. By bridging the gap between quantitative stock market data and qualitative insights from social media, we provide a holistic view of market influences, paving the way for more informed and effective stock market forecasting.

3.4.5 Extracting features from the tweets

The initial module of our predictive scheme intricately parses tweets to harvest a comprehensive set of features, spanning metadata, textual content, and author profiles. This meticulous extraction process is pivotal for encapsulating the multifaceted dimensions of social media engagement and sentiment, which are indicative of public opinion trends relevant to stock market movements.

Tweet Metadata Features: Table 1 catalogs the metadata features extracted from tweets. These features, including the designation of a tweet as a retweet or reply, the count of hashtags, URLs, and mentioned users, offer insights into the tweet's reach, engagement, and contextual relevance.

Textual Content Features: Employing advanced natural language processing (NLP) techniques, including N-grams and parts-of-speech (POS) tagging, Table 2 showcases the features derived from analyzing the textual

Table 1 Features extracted from the metadata of a tweet

IRT	Is Retweet
IRP	Is Reply To
NH	Number of Hashtags
NU	Number of URLs
NMU	Number of Mentioned Users

Table 2 Features extracted from the textual content of a tweet

TL	Length of a Tweet or Number of Words in a Tweet
NT	Number of Tokens in a tweet
NS	Number of Lemmas in the a Tweet
NL	Number of Stems

nuances of tweets. These include metrics such as tweet length, token count, and the use of lemmas and stems, crucial for understanding the semantic and sentiment layers within the tweet's content.

Author Profile Features: Reflecting the social influence and activity level of the tweet's author, Table 3 delineates features such as the total tweets, follower count, and engagement metrics. These parameters offer a window into the author's presence on Twitter, enriching the predictive model with dimensions of credibility and network impact.

The amalgamation of these diverse features from tweets into our model underlines a holistic approach to sentiment analysis, ensuring a nuanced understanding of public opinion's influence on stock market trends. By systematically analyzing tweets' metadata, textual content, and author profiles, our scheme captures the complex interplay between social media dynamics and stock market fluctuations, enhancing the precision of our predictive analytics.

3.4.6 Parts-of-speech tagging

Parts-of-speech (POS) tagging stands as a cornerstone of natural language processing (NLP), facilitating a granular analysis of textual data by assigning grammatical categories to each word. This process is crucial for dissecting the syntactic and semantic fabric of language, thereby

Table 3 Features extracted from the profile of the author of a tweet

NTw	Total Number of Tweets written by the Author
NFoll	Number of Followers
NFr	Number of Friends
NL	Number of Lists
NTU	Total Number of total URLs used by the Author
NTMU	Total Number of total Users mentioned by the Author
NTH	Total Number of total Hashtags used by the Author
NTC	Total Number of total Conversations participated by the Author
NFT	Total Number of the Author's Favorite Tweets

Table 4 Part-of-speech (POS) tags employed in the predictive scheme

CC	Conjunction	PRP$	Pronoun, Possessive
CD	Numerical, Cardinal	RB	Adverb
DT	Determiner	RBR	Adverb, Comparative
EX	Existential There	RBS	Adverb, Superlative
FW	Foreign Word	SYM	Symbol
IN	Preposition or Conjunction, Subordinating	TO	"To" as Preposition or Infinitive Marker
JJ	Adjective or Numerical, Ordinal	UH	Interjection
JJR	Adjective, Comparative	VB	Verb, Base Form
JJS	Adjective, Superlative	VBD	Verb, Past Tense
MD	Modal Auxiliary	VBG	Verb, Present Participle or Gerund
NN	Noun, Common, Singular or Mass	VBN	Verb, Past Participle
NNS	Noun, Plural	VBP	Verb, Present Tense, Not 3rd Person Singular
NNP	Noun, Proper, Singular	VBZ	Verb, Present Tense, 3rd Person Singular
NNPS	Noun, Proper, Plural	WDT	Wh-determiner
PDT	Pre-determiner	WP	Wh-pronoun
POS	Genitive Marker	WRB	Wh-adverb
PRP	Pronoun, Personal		

unlocking insights into how words collectively convey sentiment, intention, and information.

The realm of POS tagging is characterized by two predominant approaches: *rule-based* (or *linguistic*) and *stochastic* (or *probabilistic*). *Rule-based* tagging leverages a compendium of handcrafted rules to discern the grammatical function of words, taking cues from punctuation, capitalization, and the surrounding lexical context. This method hinges on the deterministic application of *context frame rules* derived from extensive linguistic research and annotated lexicons.

Conversely, *stochastic* tagging employs statistical models to infer the most probable grammatical category of a word, based on its occurrence and roles in a tagged corpus. Techniques such as Hidden Markov models (HMMs) and support vector machines (SVMs) are instrumental in this approach, offering a dynamic and context-sensitive mechanism for POS assignment [47]. These models are refined through supervised learning, utilizing corpora annotated with empirical usage patterns of words in various grammatical roles [7, 9, 12, 24, 33, 78].

Within the ambit of our predictive framework, POS tagging is applied to the textual analysis of tweets, enabling an intricate examination of linguistic structures. This analytical depth enriches the model's comprehension of public sentiment expressed on Twitter, thereby enhancing its predictive acuity regarding stock market trends. The following table (4) lists the POS tags deployed in our scheme, illustrating the spectrum of grammatical categories parsed during the analysis.

The utilization of POS tagging in our analysis signifies a methodological sophistication that elevates our ability to parse and interpret the nuanced linguistic signals embedded within tweets. By delineating the grammatical scaffolding of language, we gain a richer understanding of the sentiment and perspectives shaping public discourse on Twitter, thereby fortifying the predictive capabilities of our model with respect to stock market movements.

3.4.7 N-grams

N-gram models are pivotal in textual analysis, extracting sequences of N items (words, tokens, POS tags, characters, or symbols) from a corpus, such as tweets or sentences. These sequences, or *N-grams*, are instrumental in representing text within a vector space model, capturing the contextual continuity of language. The relevance of an N-gram in text analysis is quantified by its frequency within the corpus, which undergoes a series of preprocessing steps to ensure uniformity and relevance of the extracted features:

1. Convert all text to lowercase.
2. Eliminate 'RT' prefixes, hashtags, URLs, user mentions, emojis, and numerical values.
3. Exclude digits, punctuation, and short words (less than two characters).
4. Remove stopwords, tokenize the text, and apply stemming or lemmatization.

The mathematical foundation for N-gram probability, representing the likelihood of a sequence occurring within the corpus, is expressed as:

$$P(w_i|w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}, w_i) = \frac{c(w_{i-n+1}w_{i-n+2}\ldots w_{i-1}w_i)}{\sum_w c(w_{i-n+1}, w_{i-n+2}\ldots w_{i-1}, w_i)} \quad (1)$$

where w_i signifies the i^{th} word, with $P(*)$ and $c(*)$ denoting the probability and count of its occurrence sequence, respectively. By employing the most frequent uni-, bi-, and tri-grams extracted from tweets mentioning the company associated with the index, our model robustly approximates or predicts the future daily close prices of the U.S. stock market indices [19, 41, 62, 70, 71, 80, 85].

The employment of N-grams in our predictive framework enriches the model's understanding of language patterns within tweets, offering nuanced insights into public sentiment and discourse. This approach significantly contributes to the precision of stock market predictions, leveraging textual analysis to capture the intricacies of market-influencing conversations.

3.4.8 Stock market features

The extraction of features from daily stock market data forms a critical component of our predictive model's second module. This module gleans essential information from Yahoo! Finance, which serves as a comprehensive source for financial news and stock market dynamics. The features, detailed in Table 5, encompass a variety of daily stock market metrics for a given index in the U.S. stock market.

To enrich our analysis further, we compute two additional metrics based on daily price movements: the *High/Low Percentage* and the *Percentage Change*. These metrics provide insights into the volatility and directional movement of the market on a given day.

Table 5 Features extracted from stock market information

Number of shares
Close price
Open price
High price
Low price

The *High/Low Percentage* reflects the daily price range's relative breadth, offering a gauge of intraday volatility:

$$High/Low\ Percentage = \frac{High - Low}{Low} \times 100\% \quad (2)$$

Similarly, the *Percentage Change* measures the day's price movement relative to the opening price, indicating the market's directional trend:

$$Percentage\ Change = \frac{Open - Close}{Close} \times 100\% \quad (3)$$

These calculations, derived from the daily *High*, *Low*, *Open*, and *Close* prices, enhance our model's ability to capture and interpret significant market movements, aiding in the prediction of future stock market trends.

By integrating these stock market features with our analysis of social media sentiment, our model achieves a comprehensive view of the factors influencing stock market indices, leveraging both quantitative financial data and qualitative insights from Twitter.

3.5 Sentiment analysis and opinion mining

Sentiment analysis and opinion mining, pivotal in fields such as information retrieval and text mining, encompass a suite of methodologies designed to distill sentiments and opinions from textual data, particularly tweets. These techniques serve to categorize tweets by sentiment-positive, negative, or neutral-and by nature-subjective or objective, regarding entities such as products, events, or organizations.

The extraction of sentiment and opinion from tweets employs two primary approaches: lexicon-based methods and supervised learning algorithms. Lexicon-based approaches, such as the Vader Analyzer from Python's Natural Language Toolkit (NLTK) and the TextBlob package, leverage predefined lists of words associated with positive, negative, or neutral sentiments. Vader Analyzer, for instance, integrates various sentiment lexicons through a lexicon- and rule-based system, assigning a polarity score to each tweet that reflects its overall sentiment. This score then guides the classification of the tweet into one of the three sentiment categories.

Similarly, TextBlob utilizes lexicon- and rule-based mechanisms to assess both sentiment and subjectivity, drawing on multiple lexicons to ascertain the semantic orientation and subjectivity level of a tweet. A tweet's classification into sentiment and subjectivity categories hinges on its calculated polarity score.

On the machine learning front, techniques like Bayesian classification, support vector machine (SVM), K-nearest neighbors (k-NN), and recurrent neural networks (RNNs) offer a supervised learning approach. These methods rely on annotated datasets to train models capable of inferring the sentiment and subjectivity of tweets based on learned patterns [13, 25, 49, 64, 68, 81, 84].

Sentiment analysis and opinion mining enrich the analytical framework of our study by providing a nuanced understanding of public sentiment and opinion dynamics on Twitter. This dual approach enables a detailed exploration of how tweets reflect collective attitudes toward specific indices or entities, further informing the predictive modeling of stock market trends.

3.6 Regression techniques and deep neural networks

In order to accurately predict the daily closing prices of stock market indices, we have employed a blend of traditional regression techniques and cutting-edge deep neural network architectures. This subsection delineates the application and efficacy of these methods in our forecasting model.

Regression Techniques: The core of our predictive model incorporates several regression methodologies, notably linear regression, support vector regression (SVR), and regression trees. Each of these models plays a pivotal role in analyzing the relationship between multiple independent variables (features extracted from tweets, stock market data, and sentiment analysis) and the dependent variable (the daily close price of the index). Linear regression establishes a direct relationship, offering a clear, interpretable model, while support vector regression and regression trees cater to nonlinear data patterns, enhancing the model's adaptability to complex market dynamics.

Deep Neural Networks: To complement and augment the regression analysis, our study integrates advanced neural network architectures, specifically recurrent neural networks (RNNs) and convolutional neural networks (CNNs). RNNs are particularly adept at processing sequential data, making them ideal for analyzing time-series stock market information and tweet flows. Their structure allows them to remember and learn from past data points, capturing temporal dependencies crucial for forecasting. Conversely, CNNs are utilized for their proficiency in pattern recognition within high-dimensional data, enabling the extraction of intricate patterns and dependencies from both structured stock market data and unstructured textual data derived from tweets.

The synergistic application of these neural network models with regression techniques facilitates a comprehensive and nuanced analysis of the stock market's volatile and dynamic nature. By leveraging the strengths of both traditional and neural network models, our predictive framework achieves enhanced accuracy and robustness in

forecasting daily closing prices, accommodating the multifarious factors influencing stock market trends.

3.6.1 Linear regression

Linear regression serves as a fundamental predictive technique in our analysis, aimed at forecasting the daily closing prices of stock market indices. This method models the relationship between a dependent variable (the predictor) and one or more independent variables (features), assuming a linear correlation between them. The independent variables, denoted as $x_1, x_2, \ldots, x_{n-1}, x_n$ in our dataset, represent a variety of features derived from social media analysis, historical stock data, and sentiment scores.

In the context of our study, both single-variable (simple) and multivariable (multiple) linear regression models were employed. The choice between these two approaches hinges on the complexity of the data's underlying relationships and the number of features involved:

- **Single-variable linear regression**: It is utilized when the predictive model is based on a single feature. This approach offers a straightforward analysis of the relationship between the daily closing price and a specific, singular market or sentiment indicator.
- **Multivariable linear regression:** on the other hand, it is applied when multiple features are believed to influence the daily closing price. This model accommodates the intricate dynamics of the stock market, allowing for a comprehensive analysis that incorporates a broader array of predictive factors.

By fitting the daily closing prices (the dependent variable) against the selected features through a linear equation, we establish a predictive framework that elucidates the significant predictors of stock market movements. This methodology not only facilitates the estimation of future price trends based on current and historical data but also highlights the proportional impact of individual features on the market's direction [23, 48, 51].

The implementation of linear regression in our analysis underscores its utility in capturing and quantifying the linear associations between market indices and a diverse set of features, laying the groundwork for more sophisticated predictive models.

3.6.2 Support vector regression

Support vector regression (SVR) represents an extension of the support vector machine (SVM) framework, applied to regression problems. Unlike its classification counterpart, SVR seeks to predict a continuous outcome variable based on a set of independent variables. It operates within a high-dimensional feature space to establish a nonlinear relationship between the predictors (features) and the outcome variable (the dependent variable), which, in our case, is the daily closing price of a stock market index.

At the core of SVR is the concept of fitting the SVR function within a specified tolerance margin (ϵ), which endeavors to minimize the error of predictions while concurrently simplifying the model's complexity. This is achieved by identifying a hyperplane in the feature space that best fits the data points, allowing for a certain degree of deviation for errors within a predefined margin. The key objective of SVR is to find the optimal hyperplane that has the maximum margin, thereby ensuring that the errors for both training and unseen data are minimized.

SVR introduces flexibility through the use of kernel functions, enabling the mapping of input features into high-dimensional space where linear regression techniques can be applied to capture nonlinear relationships. Common kernels include the linear, polynomial, and radial basis function (RBF) kernels, each facilitating different types of nonlinear regression modeling.

In the context of predicting stock market trends, SVR's ability to handle complex, nonlinear relationships between multiple market indicators and stock prices makes it an invaluable tool. By leveraging past data points as samples, SVR assists in forecasting future market behaviors with a degree of accuracy that accounts for the market's inherent volatility and unpredictability [14, 30, 44].

Through the strategic application of SVR, our study harnesses the technique's robust predictive capabilities, enabling a nuanced analysis of stock market dynamics that surpasses traditional linear models. The adaptability and precision of SVR in modeling nonlinear dependencies significantly contribute to the overall efficacy of our predictive framework, enhancing our understanding of the factors driving market fluctuations.

3.6.3 Decision trees

Decision trees (DTs) represent a versatile class of non-parametric models applicable to both classification and regression tasks. Characterized by a hierarchical structure, a decision tree is initiated from a root (or parent) node and branches out into a series of internal (decision) nodes, culminating in terminal (or leaf) nodes. The initial dataset populates the root node, which then undergoes successive binary splits across the internal nodes. Each split is determined by a specific feature that best segregates the dataset, aiming to enhance the homogeneity of resultant subgroups regarding the target variable.

In the realm of regression, which is our focus, the decision process involves binary recursive partitioning – a technique that iteratively divides the dataset into increasingly smaller subsets. This division is guided by selecting

splits that minimize the sum of squared deviations within the groups, thereby refining the prediction accuracy at each step. Alternatively, splits may be chosen based on the feature contributing to the lowest impurity score, typically measured by the Gini index or mean squared error, depending on the tree's configuration.

The process progresses until it reaches a predefined condition, such as a minimum node size, at which point the node is deemed a terminal node. These leaf nodes represent the model's predictions, derived from aggregating the outcomes (e.g., mean or median of the target variable) of the data points within each terminal node.

Within our study, decision trees are employed to model the complex, nonlinear relationships between various market indicators and the daily closing price of stock market indices. The inherent flexibility of DTs to model interactions and nonlinearities without assuming a specific form for the underlying model makes them particularly suited for capturing the multifaceted dynamics of the stock market. Moreover, the interpretability of decision trees, with clear decision paths from features to outcomes, provides valuable insights into the factors most significantly impacting stock prices [43, 52, 79].

The adoption of decision trees in our predictive framework underscores their utility in distilling essential patterns and relationships from the stock market data, contributing to the robustness and accuracy of our forecasts.

3.6.4 Random forest

Random forest (RF) represents a powerful ensemble learning technique, designed to improve prediction accuracy and robustness by aggregating the outputs of multiple decision tree (DT) regressors. This method constructs a "forest" of DTs, typically numbering in the hundreds or thousands, each trained on a randomized subset of the feature space. This approach effectively decorrelates the individual trees, mitigating the overfitting tendency of single DT models and enhancing overall prediction reliability.

The operation of RF begins with the generation of diverse DTs through bagging (bootstrap aggregation), a process that introduces variability by training each tree on a bootstrapped sample of the data. This variability is further amplified by selecting a random subset of features for splitting at each node within a tree, thereby ensuring that the ensemble captures a broad range of data patterns and interactions.

The final prediction of the RF model is obtained by averaging the outputs of all individual trees. This consensus approach not only reduces the variance inherent in the predictions of individual regressors but also leverages their collective insight, often resulting in significantly improved forecasting accuracy compared to using a single DT or other non-ensemble methods.

In the context of stock market forecasting, the RF model capitalizes on the strength of multiple DTs to capture the complex, nonlinear interactions between a wide array of market indicators and sentiment analyses. By averaging the predictions of its constituent trees, RF effectively smooths out the noise and anomalies in the data, yielding more stable and reliable predictions of future market trends.

Furthermore, the intrinsic mechanism of RF to assess feature importance–based on how frequently features are used to split nodes across the forest–provides valuable insights into which variables most significantly influence stock market movements. This aspect not only enhances the predictive performance of the RF model but also contributes to our understanding of market dynamics [43, 63].

Employing RF in our study underscores the technique's adaptability and efficiency in dealing with the multifaceted nature of financial time-series data, offering a robust predictive tool for estimating future movements of stock market indices.

3.6.5 Extreme randomized trees

The extreme randomized trees (Extra tree or ET) algorithm represents a refinement of ensemble learning techniques, notably extending the principles underlying the random forest (RF) algorithm. Unlike RF, which introduces randomness through bootstrap sampling and feature subset selection at each split, the ET algorithm injects an additional layer of randomness by also randomizing the split decisions at each node of the constructed trees.

The ET algorithm generates a multitude of classification and regression trees (CARTs) using the entire dataset, diverging from RF's method of creating trees from bootstrap samples. In constructing these trees, ET diverges from the traditional approach of identifying the most discriminative thresholds for splitting features. Instead, it selects split points entirely at random for a randomly chosen subset of features at each node. This process does not seek to minimize impurity at each split actively; rather, it leverages the law of large numbers, hypothesizing that the aggregation of numerous random splits across many trees will converge to a model with desirable predictive accuracy and generalization capability.

One of the pivotal advantages of the ET algorithm lies in its ability to reduce variance without significantly increasing bias, thereby mitigating the risk of overfitting. This is achieved through the random selection of cut points, which naturally disregards outliers and diminishes the influence of highly correlated features.

The criteria for splitting nodes within an ET model are governed by two primary parameters: the number of

features considered for random selection at each split and the minimum size of samples required to continue splitting nodes. The ensemble's final prediction is derived from averaging the predictions across all trees, providing a robust estimate that reflects the consensus among a diverse set of models.

In the sphere of stock market forecasting, the ET algorithm's unique approach to ensemble learning and decision tree construction offers a compelling tool for capturing complex patterns within financial time-series data. By averaging over a large number of randomly constructed decision trees, the ET model can effectively harness the predictive power of multiple market indicators and sentiment analyses, yielding accurate and stable forecasts of future stock prices.

Employing the ET algorithm within our predictive framework capitalizes on its strengths in dealing with the intricate dynamics of the stock market, offering a significant enhancement to our model's ability to forecast daily closing prices with reduced risk of overfitting and improved generalization across varying market conditions [3, 27, 32, 54].

3.6.6 Gradient boost

The gradient boosting machine (GBM) represents a sophisticated ensemble learning technique that significantly enhances predictive modeling capabilities, particularly within the domains of classification and regression. The gradient boosting regressor (GBR) employs this approach to tackle complex datasets by sequentially combining multiple simple models, typically decision trees of limited depth (often called decision stumps), into a cohesive, powerful predictor.

Central to the GBR's methodology is the concept of boosting, where the model focuses iteratively on correcting its predecessor's errors. This process is grounded in functional gradient descent, aimed at minimizing a chosen loss function by adjusting for the residuals of preceding models. Each new tree in the sequence is built to address the residual errors of the aggregated ensemble thus far, effectively enhancing the overall model's accuracy with each iteration.

Shrinkage plays a critical role in controlling the learning rate of this iterative process, acting as a regularization mechanism that tempers the contribution of each successive tree. This deliberate slowing of the learning process helps in preventing overfitting, making the model more robust and improving its generalization capabilities on unseen data.

During the initialization phase, the GBR model parameters–such as the number of trees, the maximum number of leaves per tree, the shrinkage rate, and the maximum depth of tree interactions–are meticulously chosen. These parameters dictate the complexity and learning capacity of the model, balancing the trade-off between fitting the training data and maintaining the model's ability to generalize well.

In our study, the GBR model is adeptly applied to forecast daily closing prices of stock market indices. By leveraging a series of decision stumps as weak learners to construct a more complex and accurate composite model, GBR is particularly effective in capturing the nuanced relationships within financial time-series data. The ensemble's ability to iteratively learn from previous mistakes and adjust for them makes it exceptionally suited for dealing with the unpredictable nature of the stock market, offering precise and reliable predictions based on a broad spectrum of contributing factors.

The gradient boosting regressor's deployment within our predictive framework showcases its capacity to address the multifaceted challenges posed by stock market forecasting, benefiting from its iterative refinement approach to yield insights of high predictive value [53, 60, 77].

3.6.7 AdaBoost

AdaBoost, short for adaptive boosting, stands as a cornerstone ensemble technique in the realm of machine learning, distinguished by its ability to incrementally enhance model performance by integrating a series of weak learning algorithms. This iterative process constructs a final robust model by assigning and adjusting weight coefficients to each base learner, based on their accuracy in predicting the outcome variable.

The process begins with the deployment of an initial base learner to classify the dataset or predict the outcome variable. Subsequent learners are then sequentially introduced, each tasked with correcting the errors of the aggregate ensemble thus far. Crucially, AdaBoost adjusts the distribution of weights for the training instances, increasing the weights of those instances that were misclassified by previous models, thereby directing the learning focus of subsequent models toward these harder-to-predict cases.

After each round of prediction, AdaBoost recalibrates the weight coefficients of each base learner, emphasizing those with higher accuracy by allocating them a greater influence on the final model's predictions. The cumulative output of the AdaBoost model is a weighted sum of the predictions made by all base learners, where these weights reflect the individual learners' contribution to overall model accuracy.

In the context of forecasting stock market indices, the AdaBoost model leverages its ensemble of learners to navigate the complex and often nonlinear relationships

within financial time-series data. By progressively focusing on challenging instances–such as days with unexpected market movements–AdaBoost hones its capacity to anticipate future trends with a nuanced understanding of market dynamics.

The integration of AdaBoost into our predictive framework highlights its effectiveness in enhancing the predictive accuracy of base learners, turning a collection of simple models into a powerful composite predictor. This approach not only increases the reliability of stock market forecasts but also offers insights into the adaptive learning process's ability to cope with the financial market's inherent unpredictability [6, 10, 46].

3.6.8 eXtreme gradient boosting regression

eXtreme gradient boosting (XGBoost) stands as a cutting-edge ensemble technique that extends the conventional gradient boosting framework by introducing several optimization and regularization enhancements. Designed to tackle classification, regression, and optimization challenges, XGBoost is celebrated for its superior efficiency, scalability, and performance across diverse datasets.

At its core, XGBoost operates by sequentially building an ensemble of decision trees, each designed to correct the residuals–errors not captured by the preceding trees–in the predictions. This process involves constructing D distinct tree models, with each tree, denoted as the d^{th} tree, being trained on the residuals left by its predecessor, the $(d-1)^{th}$ tree. Through this method, each subsequent tree incrementally refines the accuracy of the ensemble, focusing on data points that are harder to predict.

XGBoost differentiates itself with several key features that enhance its tree boosting capability:

- Gradient-based optimization: XGBoost employs gradient descent to minimize loss when adding new trees, leveraging first and second derivative information of the loss function for more precise updates.
- Regularization: To combat overfitting, XGBoost integrates regularization terms into its objective function, penalizing complex models and thereby encouraging simpler, more generalizable trees.
- Shrinkage and column Subsampling: By applying shrinkage to slow down the learning in each step and randomly sampling a subset of features for splits, XGBoost further improves model robustness and prevents overfitting.
- System optimization: XGBoost optimizes computing resources, making full use of hardware capabilities, and implements advanced data structures and algorithms for speed and scalability.

In predictive modeling for stock market indices, XGBoost's ability to efficiently process large and complex datasets–incorporating a multitude of features extracted from market data and sentiment analysis–makes it an invaluable asset. By assembling an ensemble of causality-based decision trees and gradient-boosted regressors, XGBoost meticulously navigates the intricacies of financial time series, offering highly accurate forecasts that are resilient to overfitting.

The implementation of XGBoost within our framework is distinguished by its systematic approach to ensemble learning, where each model within the ensemble contributes to an increasingly refined prediction. This strategy, coupled with XGBoost's regularization and system optimizations, ensures that our predictive model achieves a delicate balance between prediction accuracy and generalization, adeptly capturing the dynamic behavior of stock market prices [42, 58, 69].

3.6.9 Recurrent neural networks

Recurrent neural networks (RNNs) are a cornerstone in the modeling of sequential data, including applications in text generation and speech recognition. Unlike traditional neural networks, RNNs process inputs in sequences, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, generating a corresponding sequence of hidden states $\mathbf{h} = (h_1, h_2, \ldots, h_n)$ that encapsulate temporal information. This unique architecture enables RNNs to retain a memory of previous inputs in the sequence, facilitating dynamic temporal behavior modeling. However, standard RNNs struggle with capturing long-term dependencies due to vanishing or exploding gradients, a challenge addressed by the development of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks.

Long Short-Term Memory (LSTM): LSTM networks incorporate gating mechanisms to mitigate the vanishing gradient problem, allowing for the preservation and selective transfer of information across the network. An LSTM unit comprises three gates:

1. *Input gate* (i_t): Determines the amount of new information to be added to the cell state.
2. *Forget gate* (f_t): Decides which information is discarded from the cell state, facilitating memory management.
3. *Output gate* (o_t): Controls the amount of information to be passed to the next hidden state from the cell state.

This gated structure ensures that LSTMs can maintain and modify their state over time, making them exceptionally suited for analyzing sequences with significant time lags between crucial events.

Gated Recurrent Unit (GRU): GRUs streamline the LSTM architecture by combining the input and forget gates into a single *update gate* (z_t) and employing a *reset gate* (r_t) to manage the transfer of information from the previous state. This simplification reduces the computational complexity while retaining the ability to model long-term dependencies. GRUs offer an efficient alternative to LSTMs, balancing performance and computational efficiency.

In the context of forecasting stock market indices, RNNs, LSTMs, and GRUs provide a sophisticated toolset for capturing the complex temporal dynamics inherent in financial time-series data. Their ability to learn from sequences of past market data and sentiment analysis allows for the prediction of future trends with enhanced accuracy. By leveraging these advanced neural network architectures, our study addresses the challenging task of modeling the sequential and temporal patterns in stock market movements, offering insights that traditional models might overlook due to their incapacity to handle long-term dependencies effectively [5, 22, 31].

3.6.10 Convolutional neural network

Convolutional neural networks (CNNs) have revolutionized the field of machine learning, particularly in applications requiring the processing of grid-like data structures, including images, videos, and time-series data. Renowned for their efficiency in high-dimensional data processing, CNNs excel in feature extraction and dimensionality reduction, which are pivotal for tasks in computer vision, natural language processing, and speech recognition. The architecture of CNNs consists of convolutional, pooling, and fully connected layers, each playing a crucial role in the network's ability to learn complex patterns:

- **Convolutional layers** perform the convolution operation, applying filters (kernels) over the input data to produce feature maps. This layer exploits local connectivity and weight sharing to efficiently capture spatial and temporal dependencies.
- **Pooling layers** reduce the spatial dimensions of the input feature maps, enhancing the network's computational efficiency and making the features robust to variations in the input data.
- **Fully connected layers** aggregate the extracted features into a flat vector, facilitating the final classification or regression output.

Convolutional Neural Network Long Short-Term Memory (CNN-LSTM): The CNN-LSTM architecture marries the spatial feature extraction capabilities of CNNs with the temporal processing power of LSTMs, forming a potent tool for analyzing sequential data with spatial characteristics. This hybrid model is particularly advantageous in financial data analytics, where the volatile and dynamic nature of the data demands sophisticated modeling techniques.

In the CNN-LSTM framework, the initial stage involves a CNN processing the input data through convolutional and flattening layers to extract and condense the spatial features. The extracted features, once flattened, are then fed into an LSTM network. The LSTM layer is adept at capturing long-term temporal dependencies, making it suitable for forecasting tasks where historical context significantly influences future trends.

- The **ReLU activation function** within the CNN layers ensures nonlinear transformations, enhancing the model's ability to learn complex spatial hierarchies in the data.
- The **flattening layer** transitions the multi-dimensional output of the CNN to a format suitable for LSTM processing, ensuring a seamless integration of spatial and temporal analysis.

The subsequent LSTM component utilizes these features to predict the dependent variable, leveraging the sequential nature of the data. The integration of CNN for feature extraction with LSTM for sequence modeling provides a comprehensive approach to predicting stock market indices, capturing both the intricate patterns in financial indicators and the temporal dynamics of market movements.

This innovative CNN-LSTM model addresses the challenges of forecasting in the financial domain by harnessing the strengths of CNNs and LSTMs, offering a nuanced understanding of market behaviors and enhancing predictive accuracy [39, 74, 75].

3.6.11 Generative adversarial network

Generative adversarial networks (GANs) represent a class of innovative generative models that have revolutionized the field of artificial intelligence, particularly in synthesizing new data instances that are indistinguishable from genuine samples. At the heart of GANs is an adversarial

Table 6 Companies, their stock indices, and top three hashtags (March 22nd-31st, 2021)

Company	Stock index	Three most frequent hashtags
Amazon	AMZN	#amazon, #deals, #sales
Apple	APPL	#Apple, #AppleMusic, #iphone
Delta	DAL	#DAL1669, #AIRBUS, #DAL1038
Google	GOOGL	#google, #DoodleForGoogle, #chrome
Microsoft	MSFT	#windows, #microsoft, #Azure

Table 7 Hyperparameters configuration for deep neural network models

Neural Network	Loss Function	Batch Size	Number of Epochs	Learning Rate
RNN	MSE	400	1,500	0.001
RNN-LSTM	MSE	400	1,500	0.001
RNN-GRU	MSE	400	1,500	0.001
CNN	MSE	400	1,500	0.001
CNN-LSTM	MSE	400	1,500	0.001
GAN	MSE	400	1,500	0.001

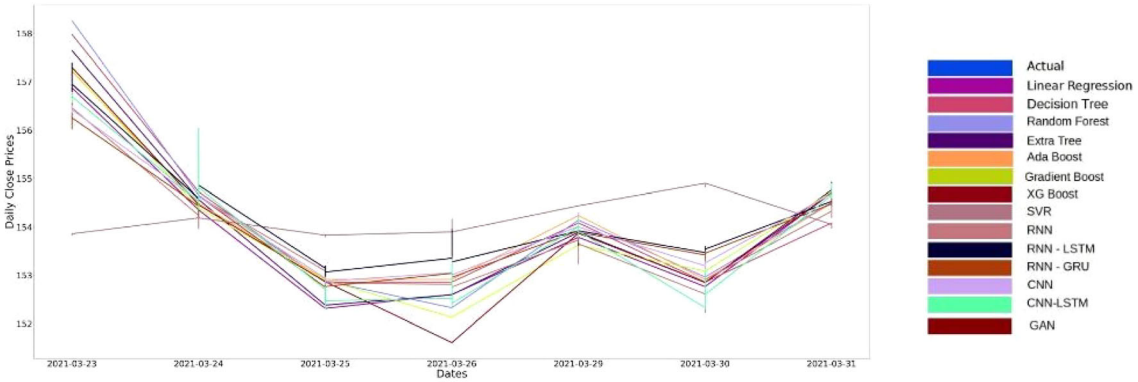

Fig. 2 Comparison of Actual and Predicted Daily Close Prices for Amazon (March 22–31st, 2021)

training process involving two distinct but interconnected models: the *generator* (G) and the *discriminator* (D).

- The **Generator** (G) is tasked with producing artificial data instances from a noise distribution or latent space $p_z(z)$. Its objective is to learn the distribution of real data so well that the generated samples are virtually indistinguishable from actual data instances. The generator does not have direct access to real data but instead learns to create data mimicking the true distribution through the adversarial process.
- The **Discriminator** (D) functions as a binary classifier, distinguishing between genuine samples from the dataset and fake samples produced by the generator. The discriminator is trained to maximize its accuracy in identifying the source of each sample, effectively guiding the generator toward producing more realistic outputs.

The adversarial training of GANs unfolds through a game-theoretic scenario where the generator strives to fool the discriminator into misclassifying its outputs as real, while the discriminator endeavors to improve its ability to discern real from fake. This dynamic competition drives both models to improve continuously, with the generator producing increasingly realistic data and the discriminator becoming more adept at detection.

In the realm of financial data analytics, GANs offer intriguing possibilities, such as generating synthetic financial datasets for training predictive models, thereby augmenting the diversity and volume of data available for analysis without compromising sensitive information. This capability is particularly valuable in scenarios where real financial data are scarce, sensitive, or imbalanced. By synthesizing realistic financial time series or indicators, GANs can provide a richer, more varied dataset for training and testing financial models, potentially unveiling novel insights into market dynamics and enhancing the robustness of predictive analytics.

4 Experimental setup and implementation

This section delineates the datasets, tools, and computational techniques utilized in our analysis for estimating and predicting the daily close prices of notable U.S. stock market indices. It also details the implementation specifics of the employed predictive scheme.

4.1 Datasets

Our analysis leveraged daily stock market data of key U.S. indices (Amazon, Apple, Delta, Google, and Microsoft) for the period from March 1st to March 31st, 2021, acquired via

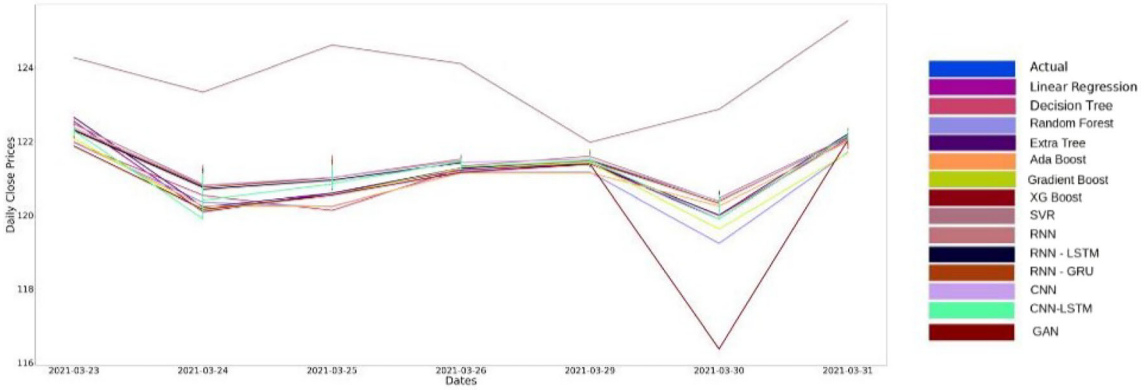

Fig. 3 Comparison of Actual and Predicted Daily Close Prices for Apple (March 22nd-31st, 2021)

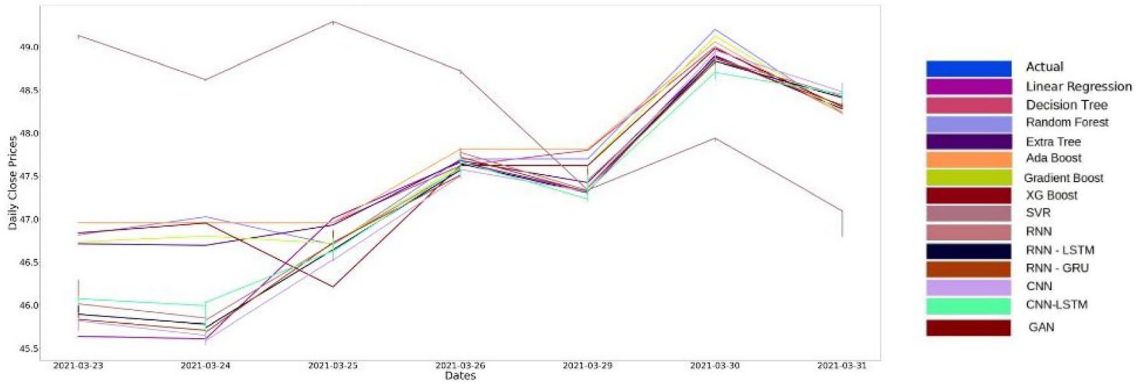

Fig. 4 Comparison of Actual and Predicted Daily Close Prices for Delta (March 22nd-31st, 2021)

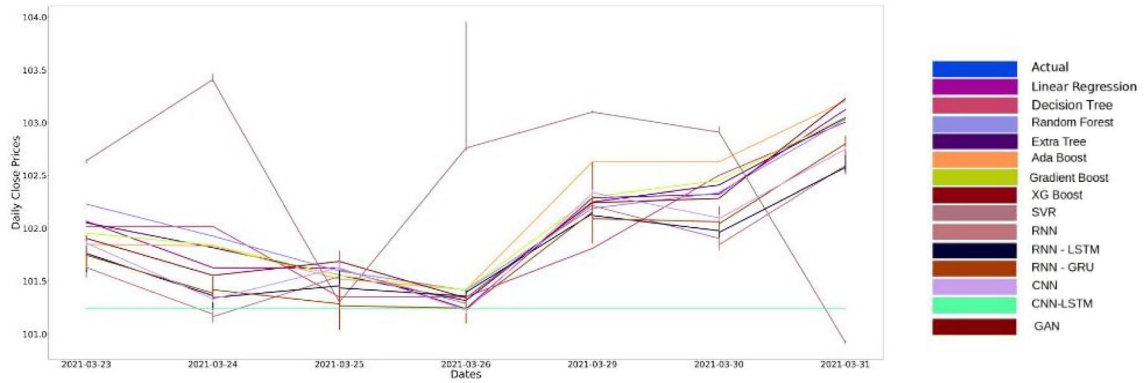

Fig. 5 Comparison of Actual and Predicted Daily Close Prices for Google (March 22nd-31st, 2021)

the Yahoo! Finance API. Given the stock market's inactivity on weekends, data corresponding to these days were omitted, particularly affecting the analysis of tweets posted during these intervals.

Distinct datasets were prepared for each company, subsequently segmented into training and estimation/prediction sets covering the periods from March 1st to March 21st and from March 22nd to March 31st, respectively, as presented in Table 6.

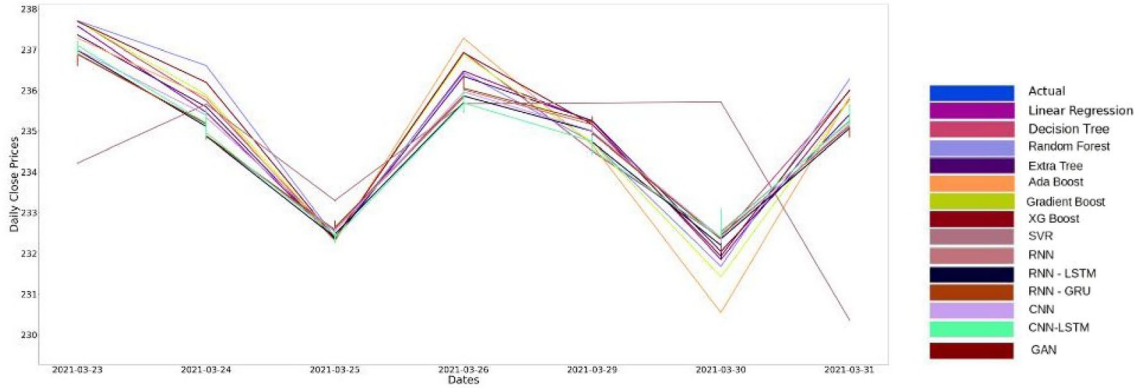

Fig. 6 Comparison of Actual and Predicted Daily Close Prices for Microsoft (March 22nd-31st, 2021)

Table 8 Social media activity metrics for listed companies (March 22nd-31st, 2021)

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

4.2 Implementation details

The scheme's implementation was executed using Python 3, leveraging a suite of packages for data acquisition, processing, and analysis:

- **Data collection** was facilitated through the "twitter" and "yfinance" packages for gathering Twitter and stock market data, respectively.
- **Textual feature extraction** employed NLTK for deriving features such as N-grams and POS tags from tweets.
- **Sentiment analysis** utilized NLTK's Vader Analyzer and TextBlob to evaluate tweets' sentiment and objectivity.
- **Machine learning models** including DNNs, implemented via "tensorflow", and the XGBoost Regressor, deployed through "xgboost", alongside various regression models configured with "scikit-learn", formed the computational backbone of our analysis.

To ensure a comprehensive evaluation, the dataset was divided into training, validation, and test sets in an 80%-20% distribution. The deep neural network models were trained with specific hyperparameters as outlined in Table 7, facilitating an optimized learning process for each model type.

This experimental setup underscores our methodological rigor in analyzing stock market trends, employing advanced computational techniques to forecast the daily closing prices of major U.S. stock indices accurately.

4.3 Model adaptability and fine-tuning

In the development of our predictive model, particular attention was paid to ensure its adaptability across various forecasting domains beyond stock market fluctuations. Central to this adaptability are several tunable parameters, specifically designed to accommodate the nuances of different datasets and forecasting objectives. For instance, the learning rate and batch size can be adjusted to balance the trade-off between training speed and model accuracy, crucial when transitioning from the volatility of stock markets to more stable datasets like consumer behavior trends. Similarly, the architecture of the neural network, including the number and size of layers, provides flexibility to model complexity varying with the domain, whether it is the nuanced predictions required for financial markets or broader trend analyses in public health.

5 Experimental results

In this section, we delve into the analysis of our predictive models' performance in estimating and forecasting the daily closing prices for the stock market indices of the companies specified in Table 6. Utilizing the comprehensive scheme described in Sect. 3, we explore how the

Table 9 Mean absolute error (MAE) analysis of model predictions compared to actual stock prices (March 22nd–31st, 2021)

Algorithm	Amazon	Apple	Delta	Google	Microsoft
LR	8.7281e−10	2.616e−09	9.5667e−11	9.5667e−11	2.9156e−10
SVR	1.2945	2.5150	1.9264	1.926	2.6089
DT	1.1560	0.9151	0.5704	0.4913	0.7086
RF	0.43949	0.3340	0.5252	0.5488	0.3533
ET	0.1123	0.0524	0.8711	0.7998	0.1204
GB	0.3195	0.3817	0.5033	0.4808	0.2777
AB	0.4346	0.8157	0.6048	0.6048	0.2500
XGB	0.2946	0.2462	0.5851	0.5148	0.5952
RNN	0.2524	0.2496	0.1970	0.1970	0.4823
LSTM	0.4856	0.2237	0.1597	0.1597	0.4068
GRU	0.3485	0.2318	0.1290	0.1290	0.4692
CNN	0.3849	0.2352	0.1704	0.1704	0.5201
CNN-LSTM	0.2038	0.0958	0.2606	0.2606	0.2956
GAN	1.2213	0.1640	0.1830	0.1830	0.3472

Table 10 Root mean square error (RMSE) analysis of model predictions compared to actual stock prices (March 22nd-31st, 2021)

Algorithm	Amazon	Apple	Delta	Google	Microsoft
LR	1.12e−09	3.20e−09	1.13E-10	6.28E-10	3.44E-10
SVR	1.5416	2.7478	2.2164	1.038	3.2810
DT	1.3793	0.3119	0.7592	0.2192	0.3604
RF	0.6837	0.4184	0.6865	0.1636	0.4824
ET	0.2697	0.0944	0.6402	0.0778	0.2185
GB	0.3673	0.4568	0.7126	0.1332	0.2998
AB	0.5671	1.6053	0.8019	0.1200	0.3327
XGB	0.3728	0.2745	1.0017	0.2246	0.6200
RNN	0.3093	0.3353	0.2301	0.3570	0.4568
LSTM	0.5508	0.2905	0.1966	0.2697	0.4924
GRU	0.3931	0.2959	0.1564	0.2678	0.4746
CNN	0.4123	0.3034	0.2270	0.1867	0.4716
CNN-LSTM	0.2385	0.1639	0.2987	0.7826	0.4877
GAN	1.2214	0.1640	0.1830	0.2718	0.0523

Table 11 Mean absolute percentage error (MAPE) analysis of model predictions compared to actual stock prices (March 22nd-31st, 2021)

Algorithm	Amazon	Apple	Delta	Google	Microsoft
LR	5.68e−10	2.16E-09	2.01E-10	5.77e−10	1.24e−10
SVR	0.8426	2.0780	4.1397	0.8821	1.1092
DT	0.6795	0.2168	1.0697	0.1618	0.1287
RF	0.3615	0.3217	1.1184	0.1423	0.1437
ET	0.1382	0.0630	0.8470	0.0568	0.0776
GB	0.1949	0.3013	1.0479	0.1197	0.0982
AB	0.2840	0.6788	1.3105	0.1080	0.1060
XGB	0.2142	0.2054	1.3743	0.2117	0.1938
RNN	0.1645	0.2071	0.4228	0.2980	0.1744
LSTM	0.3176	0.1854	0.3415	0.2488	0.1922
GRU	0.2271	0.1920	0.2757	0.2346	0.1842
CNN	0.2512	0.1949	0.3618	0.1511	0.1838
CNN-LSTM	0.1328	0.0790	0.5583	0.6510	0.1852
GAN	0.7893	0.1350	0.3868	0.2658	0.0223

Table 12 R-Squared Values for Predictive Models Across Various Companies

Algorithm	Amazon	Apple	Delta	Google	Microsoft
LR	1.0	1.0	1.0	1.0	1.0
SVR	−1.7927	−7.970	−2.593	−5.332	−1.7927
DT	0.8230	0.8760	0.3570	0.8807	0.9706
RF	0.9240	0.7691	0.6230	0.8640	0.9513
ET	0.9668	0.9990	0.7500	0.9652	0.9679
GB	0.9740	0.8846	0.6600	0.9079	0.9724
AB	0.6783	0.9125	0.5620	0.6961	0.9520
XGB	0.9713	−2.0610	0.5230	0.9154	0.9712
RNN	0.9272	0.8912	0.9630	0.6869	0.9272
LSTM	0.9500	0.8765	0.9662	0.5821	0.9500
GRU	0.9310	0.8832	0.9720	0.6742	0.9310
CNN	0.9110	0.8930	0.9523	0.7501	0.9110
CNN-LSTM	0.9660	0.9360	0.8706	0.7640	0.9656
GAN	0.9030	0.9024	0.8311	0.7018	0.9000

historical daily close prices, alongside the extracted sentiment, objectivity, and tweet-related features, influence the accuracy of our predictions.

Our evaluation hinges on a multifaceted analysis, examining not only the precision of the daily close price predictions but also the models' ability to capture the underlying market sentiments and trends as reflected through social media. By integrating sentiment analysis with traditional financial indicators, we aim to uncover the depth of correlation between public sentiment on social media and market behavior, thereby assessing the predictive power of our models in real-world scenarios.

The following subsections will present the results obtained from each neural network model outlined in the study, providing insights into their individual and collective efficacy in predicting stock market trends. Comparative analyses, along with statistical measures of performance such as mean squared error (MSE) and root mean squared error (RMSE), will be utilized to quantify the accuracy and reliability of our predictions. This comprehensive evaluation seeks to underscore the strengths and limitations of our approach, offering a critical perspective on the utility of sentiment and tweet-related features in enhancing stock market predictive analytics.

5.1 Estimating and forecasting daily close prices

The predictive models outlined in Sect. 3 were employed to estimate and forecast the daily closing prices of the stock market indices for companies listed in Table 6 from March 22nd to March 31st, 2021. This analysis period was crucial for evaluating the models' predictive accuracy under varying market conditions.

Figures 2 , 3, 4, 5 and 6 juxtapose the actual and predicted daily close prices, demonstrating the effectiveness of the applied predictive techniques:

The comparison reveals a commendable alignment between the actual and estimated prices for models such as linear regression, CNN, and CNN-LSTM, indicating their robustness in capturing the intricacies of stock market dynamics. This alignment validates the potential of employing advanced neural network architectures for financial forecasting.

Conversely, predictions from SVR and RNN models exhibited the most significant discrepancies from actual data, likely attributable to the feature set's sparsity and the recurrent issues of gradient diminishment or explosion. This condition suggests that these models, while promising, require careful feature selection and gradient management to mitigate potential overfitting.

Notably, LSTM-RNN, LSTM-GRU, GAN, and ensemble techniques displayed only marginal deviations from actual stock market closing prices, reflecting the inherently volatile, noisy, and dynamic nature of the financial markets they aim to model.

The examination of social media metrics, as detailed in Table 8, provides empirical evidence of the significant influence of Twitter activity on the forecasting accuracy of stock market prices. The data encompasses the total volume of tweets, including retweets and replies, in addition to the followers and friends of tweet authors, for the companies listed in Table 6 from March 22nd to March 31st, 2021.

Notably, Amazon and Google emerged as focal points of extensive public discourse, with 51,053 and 113,659 tweets, respectively. These figures not only reflect a high level of public engagement but also correlate with enhanced accuracy in our models' stock price predictions. For Amazon, the significant volume of tweets was paralleled by a broad diffusion through retweets (10,817) and replies (1,553), coupled with a vast network of followers (443,059,418) and friends (577,441). Similarly, Google's substantial tweet count was supported by an even larger number of retweets (87,374) and a considerable follower base (553,284,947), indicating widespread interest and interaction.

Conversely, Delta, with a comparatively lower tweet volume (3,277), retweets (1,196), and replies (196), exemplifies how lesser social media activity may correspond to larger divergences between estimated and actual stock prices. The pattern observed suggests a direct relationship between the volume and reach of company-specific discourse on Twitter and the precision of predictive models in capturing daily close prices.

This analysis underscores the pronounced effect of social media dynamics on market predictions. The extensive volume of tweets and the interaction they garner-measured through retweets, replies, and the overall engagement of the authors' networks–illustrate a robust correlation between social media activity and the dynamics of the stock market. Amazon and Google, in particular, highlight how significant public interest and engagement can enhance the predictive models' effectiveness, thereby emphasizing the critical role of social media metrics in forecasting stock market trends and volatility.

These findings offer a nuanced understanding of the interplay between market sentiment, public discourse on social media, and financial forecasting accuracy, highlighting the essential contribution of social media analytics to the predictive modeling of stock market indices.

5.2 Accuracy of the presented scheme

The efficacy of our predictive models in estimating and forecasting the daily closing prices of the indices associated with the companies listed in Table 6 is quantitatively assessed through the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). These metrics collectively offer insights into the precision of our predictions relative to actual market outcomes from March 22nd to March 31st, 2021.

MAE and RMSE provide direct measures of prediction accuracy by quantifying the average magnitude and squared deviations of errors between predicted and actual values, respectively. MAPE extends this analysis to a relative scale, offering a percentage-based assessment of

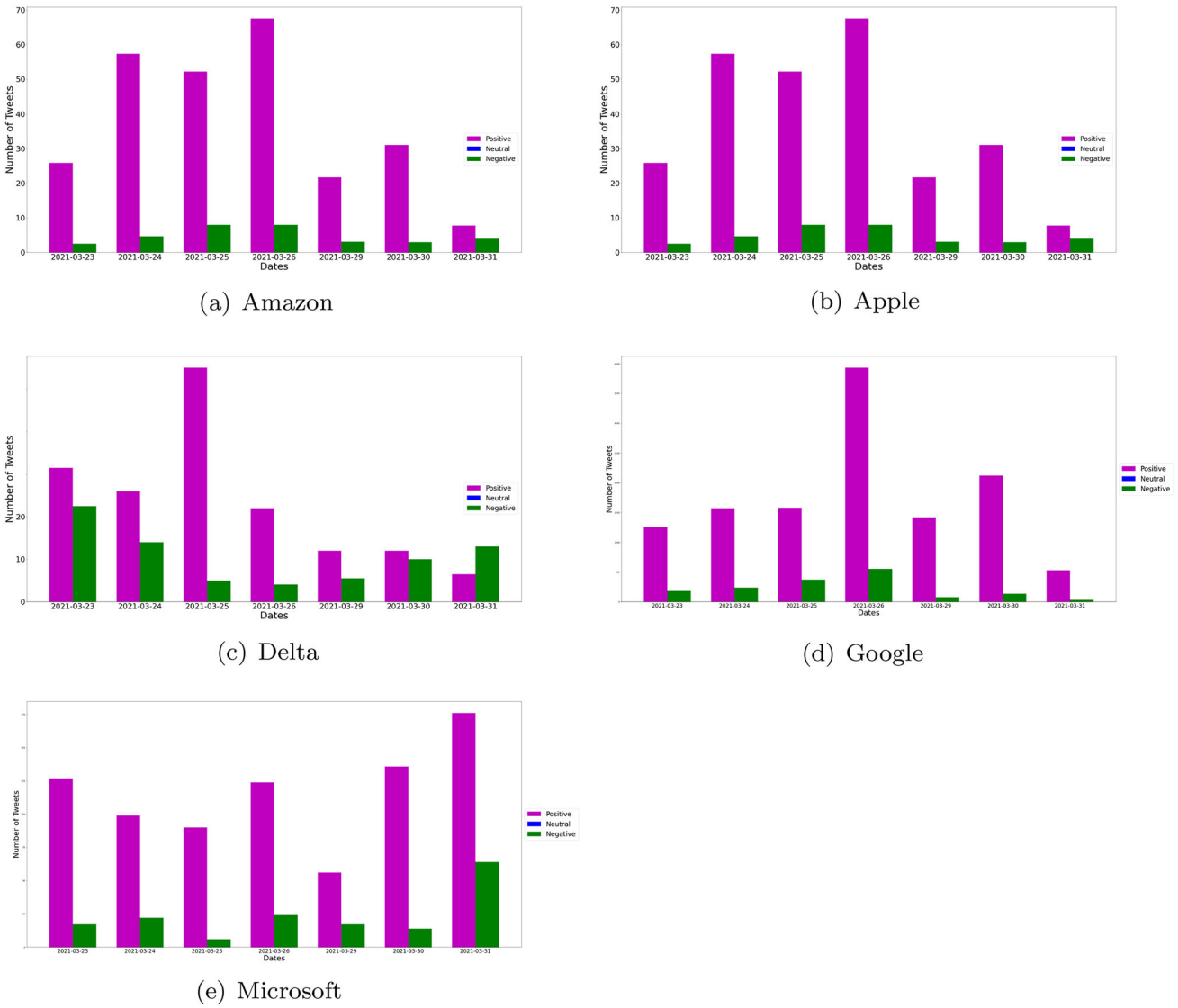

Fig. 7 Distribution of Sentiment Across Tweets Mentioning Selected Companies (March 22nd-31st, 2021)

Table 13 Sentiment classification of tweets mentioning listed companies (March 22nd-31st, 2021)

Date	Amazon	Apple	Delta	Google	Microsoft
22-03-2021	Positive	Positive	Positive	Positive	Positive
23-03-2021	Positive	Positive	Positive	Positive	Positive
24-03-2021	Positive	Positive	Positive	Positive	Positive
25-03-2021	Positive	Positive	Positive	Positive	Positive
25-03-2021	Positive	Positive	Positive	Positive	Positive
26-03-2021	Positive	Positive	Positive	Positive	Positive
27-03-2021	Positive	Positive	Positive	Positive	Positive
28-03-2021	Positive	Positive	Positive	Positive	Positive
29-03-2021	Positive	Positive	Positive	Positive	Positive
30-03-2021	Positive	Positive	Positive	Positive	Positive
31-03-2021	Positive	Positive	Negative	Positive	Positive

prediction accuracy, thus facilitating a comparison across different stock indices irrespective of their price scales.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y_i}| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y_i})^2} \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=0}^{N}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100 \tag{6}$$

where N represents the total number of prediction samples, y_i denotes the actual close price, and $\hat{y_i}$ signifies the estimated close price on the i^{th} day.

Upon examining the MAE, RMSE, and MAPE values across different models and companies (Tables 9, 10 and 11), a nuanced understanding of model performance emerges. Linear regression (LR) exhibited remarkable precision, especially for Amazon, indicating minimal deviation in predictions. However, the support vector regression (SVR) and recurrent neural network (RNN) models showed significant error rates, suggesting potential limitations in their capacity to model the stock market's complex dynamics accurately.

Interestingly, the MAPE metric illuminated the percentage discrepancies between predicted and actual prices, revealing the proportional impact of predictive inaccuracies. This metric was particularly insightful for assessing models like CNN-LSTM, which demonstrated superior performance by minimizing percentage errors, highlighting its efficacy in closely tracking market movements.

The comparative analysis across algorithms reveals that no single model consistently outperforms others across all companies, underscoring the diverse nature of stock market data and the varying degrees of model sensitivity to data sparsity and noise. Models like gradient boosting (GB) and AdaBoost (AB) presented low error rates across several companies, showcasing their robustness. Conversely, the generative adversarial network (GAN) exhibited higher errors for specific companies like Amazon and Delta,

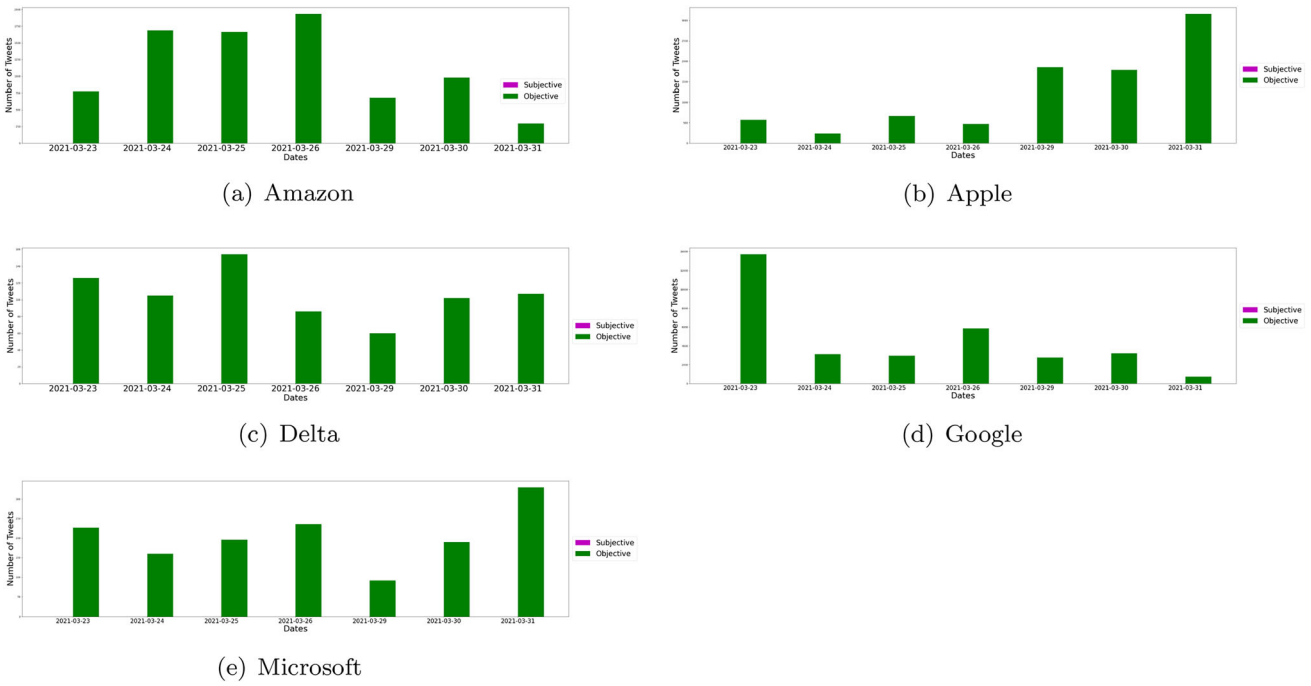

Fig. 8 Distribution of Objectivity and Subjectivity Across Tweets Mentioning Selected Companies (March 22nd-31st, 2021)

Table 14 Objectivity classification of tweets mentioning listed companies (March 22nd-31st, 2021)

Date	Amazon	Apple	Delta	Google	Microsoft
22-03-2021	Objective	Objective	Objective	Objective	Objective
23-03-2021	Objective	Objective	Objective	Objective	Objective
24-03-2021	Objective	Objective	Objective	Objective	Objective
25-03-2021	Objective	Objective	Objective	Objective	Objective
25-03-2021	Objective	Objective	Objective	Objective	Objective
26-03-2021	Objective	Objective	Objective	Objective	Objective
27-03-2021	Objective	Objective	Objective	Objective	Objective
28-03-2021	Objective	Objective	Objective	Objective	Objective
29-03-2021	Objective	Objective	Objective	Objective	Objective
30-03-2021	Objective	Objective	Objective	Objective	Objective
31-03-2021	Objective	Objective	Objective	Objective	Objective

indicating possible overfitting or inadequacies in capturing market volatilities.

This evaluation underscores the complexity of financial market forecasting, where data sparsity and external market factors can significantly influence model accuracy. The findings advocate for a multifaceted approach to model selection and the importance of considering a range of error metrics for comprehensive performance assessment. Furthermore, the superior insights provided by MAPE in reflecting actual market fluctuations reinforce its value as a critical metric in the evaluation framework for predictive models.

Overall, our analysis affirms the potential of advanced machine learning techniques in financial predictions while highlighting the necessity for meticulous model selection and evaluation strategies to optimize forecasting accuracy.

5.3 Twitter sentiment analysis, tweet diffusion, and stock market volatility

The sentiment and subjectivity of tweets mentioning the companies listed in Table 6 were analyzed to understand their impact on stock market volatility. By calculating the polarity scores, tweets were categorized as positive, negative, or neutral, further distinguishing them as objective or subjective. This analysis provides insights into the emotional tone and objectivity of public discourse surrounding these companies.

5.3.1 Sentiment analysis and stock market impact

Sentiment analysis categorizes the emotional tone of tweets, which is pivotal in assessing how public sentiment reflected through Twitter affects stock market dynamics. The overall sentiment distribution and its correlation with stock market movements are depicted in Fig. 7 and quantitatively summarized in Table 13.

The diffusion of tweets, detailed in Table 8, highlights the spread of sentiment across the Twitter network, potentially influencing the stock market. Positive sentiment generally correlates with increased buying activity, potentially leading to price surges, while negative sentiment can trigger sell-offs, resulting in price drops.

5.3.2 Subjectivity, objectivity, and market stability

The analysis extends to evaluating the objectivity and subjectivity of tweets, understanding how factual versus opinion-based tweets impact stock prices. The distribution of tweet objectivity is visualized in Fig. 8 and detailed in Table 14.

Objective tweets tend to stabilize market sentiment by providing factual information, whereas subjective tweets, particularly negative ones, can significantly alter market perceptions and lead to volatility. The rapid spread of subjective content can amplify market reactions, underscoring the impact of social media on financial markets.

5.4 Discussion

This analysis reveals a complex interplay between Twitter sentiment, subjectivity, and stock market volatility, highlighting how social media discourse can influence investor behavior and market trends. The widespread positive sentiment tends to boost market optimism, leading to potential overvaluations, whereas negative sentiment can incite market corrections or downturns. Furthermore, the balance between objective information and subjective opinions on social media platforms plays a crucial role in shaping market stability and investor confidence.

Overall, the findings emphasize the significance of sentiment analysis in monitoring market sentiment and predicting stock market movements, providing valuable insights for investors, analysts, and financial institutions

aiming to navigate the intricacies of market dynamics influenced by social media.

5.5 Implications of model flexibility

Our model's inherent flexibility not only demonstrates its efficacy in forecasting stock market prices based on social media sentiment but also holds promise for a range of other applications. Hypothetically, adapting our model to forecast election outcomes would entail recalibrating input features to incorporate political sentiment indicators from social media platforms. Such an adaptation might require fine-tuning the dropout rate to mitigate overfitting risks associated with the highly polarized nature of political discourse. Furthermore, adjustments to the optimization algorithm could enhance the model's responsiveness to the dynamic and rapidly evolving landscape of political campaigns. These examples underscore the model's potential utility across various domains, highlighting its capacity to contribute valuable insights into diverse phenomena.

6 Conclusions and future work

This paper introduced a predictive scheme for forecasting the daily close prices of various U.S. stock market indices. Leveraging the correlation between historical stock prices, Twitter-derived public sentiment, and opinion, the model integrated regression techniques with deep neural networks (DNNs) to analyze features from tweet metadata, textual content, and authors' social media presence.

Evaluation through mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) demonstrated the effectiveness of CNN-LSTM, extra tree, and linear regression models, underscoring their capacity to handle complex datasets while maintaining low error rates. These findings affirm the viability of hybrid and ensemble approaches in capturing the nuanced relationship between stock prices and social sentiment. However, challenges such as overfitting, highlighted by the performance of support vector regression (SVR) and regression trees, underscore the necessity for meticulous feature selection and model training.

The significant impact of public sentiment and opinions on stock price predictions suggests the potential of social media analytics as a valuable tool in financial forecasting. The inclusion of diverse Twitter data, including user engagement metrics and textual analysis, enriches the model's predictive accuracy, offering a comprehensive view of market sentiment dynamics.

Future research could extend beyond the confines of this study to explore the prediction of stock market volatility using real-time data from various social media platforms and search engines. This approach would benefit from the deployment of advanced DNN architectures or hybrid models tailored for sentiment and feature extraction from large-scale social media data. Specifically, the exploration of large language models for nuanced sentiment analysis and feature extraction from tweets holds promise for enhancing forecasting accuracy.

Moreover, addressing the challenges posed by dynamic optimization in financial markets, future work could aim to streamline the prediction process, making it more accessible and effective for real-time financial decision making. Investigations into simplifying dynamic optimization problems [1, 2, 17, 20, 28, 29] could provide a framework for more efficient and scalable predictive models in financial technology applications.

The convergence of social media analytics, advanced computational models, and financial market forecasting represents a fertile ground for innovative research, with the potential to significantly impact both academic study and industry practice in financial analysis and decision making.

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

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Abouelmagd EI, Awad ME, Elzayat EMA, Abbas IA (2014) Reduction the secular solution to periodic solution in the generalized restricted three-body problem. Astrophys Space Sci 350:495–505
2. Agushaka JO, Ezugwu AE, Abualigah L (2023) Gazelle optimization algorithm: a novel nature-inspired metaheuristic optimizer. Neural Comput Appl 35(5):4099–4131
3. Ahmad MW, Mourshed M, Rezgui Y (2018) Tree-based ensemble methods for predicting pv power generation and their comparison with support vector regression. Energy 164:465–474
4. Altay E, Satman MH (2005) Stock market forecasting: artificial neural network and linear regression comparison in an emerging market. J Financial Manag Anal 18(2):18
5. Althelaya KA, El-Alfy ESM, Mohammed S (2018) Stock market forecast using multivariate analysis with bidirectional and stacked (lstm, gru). In: 21st Saudi computer society national computer conference (NCC). pp. 1–7. IEEE
6. Ampomah Ernest Kwame, Qin Zhiguang, Nyame Gabriel, Botchey Francis Effirm (2021) Stock market decision support modeling with tree-based AdaBoost ensemble machine learning models. Informatica. https://doi.org/10.31449/inf.v44i4.3159

7. Antony PJ, Soman KP (2011) Parts of speech tagging for indian languages: a literature survey. Int J Comput Appl 34(8):0975–8887
8. Baltas A, Kanavos A, Tsakalidis AK (2016) An apache spark implementation for sentiment analysis on twitter data. In: 2nd International workshop algorithmic aspects of cloud computing (ALGOCLOUD). Lecture notes in computer science, vol. 10230, pp. 15–25
9. Banko M, Moore RC (2004) Part-of-speech tagging in context. In: 20th International conference on computational linguistics (COLING)
10. Barrow DK, Crone SF (2016) A comparison of adaboost algorithms for time series forecast combination. Int J Forecast 32(4):1103–1119
11. Bhuriya D, Kaushal G, Sharma A, Singh U (2017) Stock market predication using a linear regression. In: International conference of electronics, communication and aerospace technology (ICECA). vol. 2, pp. 510–513. IEEE
12. Binulal GS, Goud PA, Soman KP (2009) A svm based approach to telugu parts of speech tagging using svmtool. Int J Recent Trends Eng 1(2):183
13. 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
14. Chahboun S, Maaroufi M (2021) Performance comparison of support vector regression, random forest and multiple linear regression to forecast the power of photovoltaic panels. In: 9th International renewable and sustainable energy conference (IRSEC). pp. 1–4. IEEE
15. Chen S, He H (2018) Stock prediction using convolutional neural network. In: IOP Conference series: materials science and engineering. vol. 435, p. 012026
16. Drakopoulos G, Kanavos A, Mylonas P, Sioutas S (2021) Discovering sentiment potential in twitter conversations with hilbert-huang spectrum. Evol Syst 12(1):3–17
17. Du B, Liu Y, Abbas IA (2016) Existence and asymptotic behavior results of periodic solution for discrete-time neutral-type neural networks. J Franklin Inst 353(2):448–461
18. Efendi R, Arbaiy N, Deris MM (2018) A new procedure in stock market forecasting based on fuzzy random auto-regression time series model. Inf Sci 441:113–132
19. Egghe L (2000) The distribution of n-grams. Scientometrics 47(2):237–252
20. Ezugwu AE, Agushaka JO, Abualigah L, Mirjalili S, Gandomi AH (2022) Prairie dog optimization algorithm. Neural Comput Appl 34(22):20017–20065
21. Feng Y, Wang S (2017) A forecast for bicycle rental demand based on random forests and multiple linear regression. In: 16th IEEE/ACIS International conference on computer and information science (ICIS). pp. 101–105
22. Fu R, Zhang Z, Li L (2016) Using lstm and gru neural network methods for traffic flow prediction. In: 31st youth academic annual conference of chinese association of automation (YAC). pp. 324–328. IEEE
23. Fumo N, Biswas MAR (2015) Regression analysis for prediction of residential energy consumption. Renew Sustain Energy Rev 47:332–343
24. Gashaw I, Shashirekha HL (2020) Machine learning approaches for amharic parts-of-speech tagging. CoRR **abs/2001.03324**
25. Gujjar JP, Kumar HP (2021) Sentiment analysis: textblob for decision making. Int J Sci Res Eng Trends 7(2):1097–1099
26. Gupta R, Chen M (2020) Sentiment analysis for stock price prediction. In: 3rd Conference on multimedia information processing and retrieval (MIPR). pp. 213–218. IEEE
27. Hameed MM, Alomar MK, Khaleel F, Al-Ansari N (2021) An extra tree regression model for discharge coefficient prediction: novel, practical applications in the hydraulic sector and future research directions. Math Probl Eng 2021:1–19
28. Hu G, Guo Y, Wei G, Abualigah L (2023) Genghis khan shark optimizer: a novel nature-inspired algorithm for engineering optimization. Adv Eng Inform 58:102210
29. Hu G, Zheng Y, Abualigah L, Hussien AG (2023) DETDO: an adaptive hybrid dandelion optimizer for engineering optimization. Adv Eng Inform 57:102004
30. Hu J, Gao P, Yao Y, Xie X (2014) Traffic flow forecasting with particle swarm optimization and support vector regression. In: 17th international conference on intelligent transportation systems (ITSC). pp. 2267–2268. IEEE
31. Huynh HD, Dang LM, Duong D (2017) A new model for stock price movements prediction using deep neural network. In: 8th international symposium on information and communication technology. pp. 57–62. ACM
32. Joh V, Liu Z, Guo C, Mita S, Kidono K (2016) Real-time lane estimation using deep features and extra trees regression. In: 7th Pacific-Rim symposium on image and video technology (PSIVT). pp. 721–733
33. Kanakaraddi SG, Nandyal SS (2018) Survey on parts of speech tagger techniques. In: International conference on current trends towards converging technologies (ICCTCT). pp. 1–6. IEEE
34. Kanavos A, Kafeza E, Makris C (2015) Can we rank emotions? A brand love ranking system for emotional terms. In: IEEE international congress on big data. pp. 71–78. IEEE Computer Society
35. Kanavos A, Nodarakis N, Sioutas S, Tsakalidis AK, Tsolis D, Tzimas G (2017) Large scale implementations for twitter sentiment classification. Algorithms 10(1):33
36. Kanavos A, Perikos I, Hatzilygeroudis I, Tsakalidis AK (2018) Emotional community detection in social networks. Comput Electr Eng 65:449–460
37. Kanavos A, Vonitsanos G, Mohasseb A, Mylonas P (2020) An entropy-based evaluation for sentiment analysis of stock market prices using twitter data. In: 15th International workshop on semantic and social media adaptation and personalization (SMAP). pp. 1–7. IEEE
38. Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK (2013) Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Appl Soft Comput 13(2):947–958
39. Koprinska I, Wu D, Wang Z (2018) Convolutional neural networks for energy time series forecasting. In: International joint conference on neural networks (IJCNN). pp. 1–8. IEEE
40. Kumar M, Thenmozhi M (2006) Forecasting stock index movement: A comparison of support vector machines and random forest. In: Indian institute of capital markets 9th capital markets conference paper
41. Li A, Jabri A, Joulin A, van der Maaten L (2017) Learning visual n-grams from web data. In: International conference on computer vision (ICCV). pp. 4193–4202. IEEE
42. Li W, Yin Y, Quan X, Zhang H (2019) Gene expression value prediction based on xgboost algorithm. Front Genet 10:1077
43. Li Y, Zou C, Berecibar M, Nanini-Maury E, Chan JCW, van den Bossche P, Mierlo JV, Omar N (2018) Random forest regression for online capacity estimation of lithium-ion batteries. Appl Energy 232:197–210
44. Lin K, Lin Q, Zhou C, Yao J (2007) Time series prediction based on linear regression and SVR. In: 3rd International conference on natural computation (ICNC). pp. 688–691. IEEE
45. Liu H, Long Z (2020) An improved deep learning model for predicting stock market price time series. Digit Signal Process 102:102741
46. Liu Q, Wang X, Huang X, Yin X (2020) Prediction model of rock mass class using classification and regression tree integrated adaboost algorithm based on tbm driving data. Tunn Undergr Space Technol 106:103595

47. Liu Y, Liu W, Obaid MA, Abbas IA (2016) Exponential stability of markovian jumping cohen-grossberg neural networks with mixed mode-dependent time-delays. Neurocomputing 177:409–415
48. Maulud DH, Abdulazeez AM (2020) A review on linear regression comprehensive in machine learning. J Appl Sci Technol Trends 1(4):140–147
49. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
50. Mehtab S, Sen J (2020) Stock price prediction using convolutional neural networks on a multivariate timeseries. CoRR **abs/2001.09769**
51. Montgomery DC, Peck EA, Vining GG (2021) Introduction to linear regression analysis. John Wiley & Sons
52. Nabipour M, Nayyeri P, Jabani H, Mosavi A, Salwana E, Shamshirband S (2020) Deep learning for stock market prediction. Entropy 22(8):840
53. Nasiboglu R, Nasibov EN (2023) WABL method as a universal defuzzifier in the fuzzy gradient boosting regression model. Expert Syst Appl 212:118771
54. Okoro EE, Obomanu T, Sanni SE, Olatunji DI, Igbinedion P (2022) Application of artificial intelligence in predicting the dynamics of bottom hole pressure for under-balanced drilling: extra tree compared with feed forward neural network model. Petroleum 8(2):227–236
55. Oliveira N, Cortez P, Areal N (2013) Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In: 3rd international conference on web intelligence, mining and semantics (WIMS). p. 31. ACM
56. Pang XW, Zhou Y, Wang P, Lin W, Chang V (2020) An innovative neural network approach for stock market prediction. J Supercomput 76(3):2098–2118
57. Peng Z, Huang Q, Han Y (2019) Model research on forecast of second-hand house price in chengdu based on xgboost algorithm. In: 11th International conference on advanced infocomm technology (ICAIT). pp. 168–172. IEEE
58. Pesantez-Narvaez J, Guillen M, Alcañiz M (2019) Predicting motor insurance claims using telematics data-xgboost versus logistic regression. Risks 7(2):70
59. Polamuri SR, Srinivas K, Mohan AK (2019) Stock market prices prediction using random forest and extra tree regression. Int J Recent Technol Eng (IJRTE) 8(1):1224–1228
60. Ponraj AS, Vigneswaran T (2020) Daily evapotranspiration prediction using gradient boost regression model for irrigation planning. J Supercomput 76(8):5732–5744
61. Qiu Y, Yang H, Lu S, Chen W (2020) A novel hybrid model based on recurrent neural networks for stock market timing. Soft Comput 24(20):15273–15290
62. Robertson AM, Willett P (1998) Applications of n-grams in textual information systems. J Documentation 54(1):48–67
63. Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev 71:804–818
64. Sahayak V, Shete V, Pathan A (2015) Sentiment analysis on twitter data. Int J Innov Res Adv Eng (IJIRAE) 2(1):178–183
65. Saravanos C, Kanavos A (2023) Forecasting stock market alternations using social media sentiment analysis and deep neural networks. In: 14th International conference on information, intelligence, systems & applications (IISA). pp. 1–8. IEEE
66. Saravanos C, Kanavos A (2023) Forecasting stock market alternations using social media sentiment analysis and regression techniques. In: International conference on artificial intelligence applications and innovations (AIAI). IFIP advances in information and communication technology, vol. 677, pp. 335–346. Springer
67. Sharma N, Juneja A (2017) Combining of random forest estimates using lsboost for stock market index prediction. In: 2nd international conference for convergence in technology (I2CT). pp. 1199–1202. IEEE
68. Sharma V, Khemnar R, Kumari R, Mohan BR (2019) Time series with sentiment analysis for stock price prediction. In: 2nd International conference on intelligent communication and computational techniques (ICCT). pp. 178–181. IEEE
69. Shehadeh A, Alshboul O, Mamlook REA, Hamedat O (2021) Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, lightgbm, and xgboost regression. Autom Constr 129:103827
70. Sidorov G, Velasquez F, Stamatatos E, Gelbukh AF, Chanona-Hernández L (2012) Syntactic dependency-based n-grams as classification features. In: 11th Mexican international conference on artificial intelligence (MICAI). vol. 7630, pp. 1–11
71. Sidorov G, Velasquez F, Stamatatos E, Gelbukh AF, Chanona-Hernández L (2014) Syntactic n-grams as machine learning features for natural language processing. Expert Syst Appl 41(3):853–860
72. Staffini A (2022) Stock price forecasting by a deep convolutional generative adversarial network. Front Artif Intell 5:837596
73. Takahashi S, Chen Y, Tanaka-Ishii K (2019) Modeling financial time-series with generative adversarial networks. Phys A 527:121261
74. Tian C, Ma J, Zhang C, Zhan P (2018) A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network. Energies 11(12):3493
75. Tsantekidis A, Passalis N, Tefas A, Kanniainen J, Gabbouj M, Iosifidis A (2017) Forecasting stock prices from the limit order book using convolutional neural networks. In: 19th IEEE conference on business informatics (CBI). pp. 7–12. IEEE
76. Vernikou S, Lyras A, Kanavos A (2022) Multiclass sentiment analysis on covid-19-related tweets using deep learning models. Neural Comput Appl 34(22):19615–19627
77. Wang F, Mamo T (2020) Gradient boosted regression model for the degradation analysis of prismatic cells. Comput Indus Eng 144:106494
78. Wegari GM, Meshesha M (2011) Parts of speech tagging for afaan oromo. Int J Adv Comput Sci Appl 1(3):1–5
79. Xu M, Watanachaturaporn P, Varshney PK, Arora MK (2005) Decision tree regression for soft classification of remote sensing data. Remote Sens Environ 97(3):322–336
80. Yadav N, Joglekar H, Rao RPN, Vahia MN, Adhikari R, Mahadevan I (2010) Statistical analysis of the indus script using n-grams. PLoS ONE 5(3):e9506
81. Yao J (2019) Automated sentiment analysis of text data with nltk. In: Journal of physics: conference series. vol. 1187, p. 052020
82. Ye ZJ, Schuller BW (2021) Capturing dynamics of post-earnings-announcement drift using a genetic algorithm-optimized xgboost. Expert Syst Appl 177:114892
83. Yoshihara A, Fujikawa K, Seki K, Uehara K (2014) Predicting stock market trends by recurrent deep neural networks. In: 13th Pacific Rim international conference on artificial intelligence (PRICAI). Lecture notes in computer science, vol. 8862, pp. 759–769. Springer
84. Zhang Lei, Wang Shuai, Liu Bing (2018) Deep learning for sentiment analysis: a survey. WIREs Data Mining Knowl Dis. https://doi.org/10.1002/widm.1253
85. Zhang S, Dong N (2003) An effective combination of different order n-grams. In: 17th Pacific Asia conference on language, information and computation (PACLIC). pp. 251–256

86. Zhou X, Pan Z, Hu G, Tang S, Zhao C (2018) Stock market prediction on high-frequency data using generative adversarial nets. Mathematical problems in engineering