Received: 4 November 2023 | Revised: 13 March 2024 | Accepted: 1 April 2024 | Published online: 8 April 2024

RESEARCH ARTICLE

Journal of Comprehensive Business Administration Research
2024, Vol. 00(00) 1–12
DOI: 10.47852/bonviewJCBAR42022006

BON VIEW PUBLISHING

Impact of Social Media Sentiments on Stock Market Behavior: A Machine Learning Approach to Analyzing Market Dynamics

Theeshanthani Kandasamy[1] and Kamal Bechkoum[1,*]

[1]*Computing School, University of Gloucestershire, United Kingdom*

Abstract: Social media has become a valuable tool for informed decision-making. This research delves into the influence of Twitter sentiments on the stock market's movements and price fluctuations, specifically focusing on Tesla Inc. and the tweets of Elon Musk. A combination of deductive and inductive reasoning approaches is used to explore the intricate relationship between the social media platform and the stock market. Methodologically, the Twitter data undergoes rigorous processing to derive features for the machine learning predictive model, and the sentiments are extracted using the Valence Aware Dictionary and Sentiment Reasoner tool. This study emphasizes the usefulness of social media in predictive modeling while underscoring the importance of evaluating data reliability considering challenges such as spam tweets and geographical relevance. Multiple machine learning models are tested against four distinct datasets addressing the high stock price volatility. XG Boost and Random Forest Regressor emerge as the most effective performers, particularly when moving averages are included, showing enhanced performance. This research establishes an evident correlation between social media sentiments and stock market movements, however with limited predicting power. It is also noted that integrating traditional financial metrics enriches the understanding of stock market dynamics while enhancing the model's predictability.

Keywords: social media, stock market, sentiment analysis, moving averages, stock market behavior

1. Introduction

In recent years, social media has witnessed a spectacular surge in popularity, establishing itself as a widespread and important communication tool. It has transitioned from a virtual community into a thriving marketplace and an essential marketing tool due to its wide reach, significant effect, and strong influence. Globally, 4.9 billion individuals use social media, and this figure is expected to rise to 5.85 billion by 2027 [1]. The increased internet and social media usage has led to an unprecedented proliferation of individual viewpoints on various topics, and these viewpoints often contain valuable sentiments suitable for sentiment analysis [2].

Simultaneously, technological advancements in recent decades have contributed significantly to the expansion of the stock market. Investors, in their pursuit of optimized returns and risk mitigation, continually seek novel strategies and approaches [3]. Research related to stock market prediction has attracted significant scholarly attention recently due to its rapid growth, unique challenges, and technological advancements. Financial data, on the other hand, is known for its noise and complexity, making stock market price prediction very challenging [4]. Furthermore, stock markets are greatly influenced by publicly available information from social media, news sources, and financial reports. Consequently, stock market predictions based only on historical financial data could be misleading, encouraging scholars to explore external factors that affect the stock market [5].

Since the 1960s, the efficient market hypothesis (EMH) has been a fundamental concept in modern finance [6, 7]. According to Eugene Fama [7], the EMH is when a market is fully efficient and market prices reflect all available information, making it impossible for investors to gain capital gains from price fluctuations [8–11]. Despite the longstanding dominance of the EMH, skepticism has risen regarding its absolute efficiency [12]. This study delves into market inefficiencies particularly driven by investor moods and sentiments, exploring the potential of leveraging such inefficiencies for stock market prediction.

This research holds significance in highlighting the intricate relationship between social media sentiments, specifically on Twitter, and stock market dynamics with a focus on Tesla Inc. and Elon Musk's tweets. As social media emerges as a pivotal medium for sharing knowledge and expression of public opinion, understanding its impact on financial markets becomes imperative. Through the use of advanced sentiment analysis and machine learning models, the research challenges the traditional EMH by identifying potential market inefficiencies influenced by investor sentiments. The use of moving averages in predictive models emphasizes the importance of combining traditional financial measures with social media data to achieve greater accuracy.

Furthermore, this study systematically investigates the impact of incorporating tweets from a renowned influencer, as demonstrated by Elon Musk's tweets, on stock market dynamics and the predictive accuracy of the algorithms used. The study

*Corresponding author: Kamal Bechkoum, Computing School, University of Gloucestershire, United Kingdom. Email: kbechkoum@glos.ac.uk

explicitly studies a prominent business figure's social media sentiment to assess the extent to which such high-profile persons can amplify or modify the generally accepted correlation between social media sentiment and stock market behavior. Beyond theoretical contributions, the study has practical consequences for investors, analysts, and policymakers, providing detailed insights into the complex role that social media sentiments play in driving stock market movements.

The primary objective of this study is to evaluate the effectiveness of the machine learning models in predicting stock market behavior using sentiments from Twitter, addressing key research questions on effectiveness, accuracy, and the influence of influential figures while recognizing the complexities present in social media data, including a vast amount of irrelevant information [13]. This study addresses the following research questions (RQs):

- RQ1. Can the sentiments extracted from social media (tweets) be used to predict stock prices effectively?
- RQ2. Which machine learning model, when trained and tested with sentiment scores and a financial indicator, produces the most accurate prediction?
- RQ3. How does including tweets from well-known figures, such as Elon Musk, impact the correlation between social media sentiments and stock market behavior, and what is its influence on predictive model efficacy?
- RQ4. How does including moving averages into predictive models improve the accuracy of stock market predictions when using social media sentiments.

The Valence Aware Dictionary and Sentiment Reasoner (VADER) tool is employed to generate sentiment scores from Twitter data. This study uses this tool to analyze individual sentiments expressed in tweets. Six diverse machine learning models, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) Regressor, Support Vector Regressor (SVR), Extreme Gradient Boosting (XG Boost), and Prophet, are harnessed to predict stock market prices. The sentiments generated serve as the primary feature of the model complemented by the moving average as the secondary feature. This research represents a comprehensive exploration of the interplay between Twitter sentiments and stock market prices, providing insights into the potential predictive power of social media sentiment analysis in the financial domain.

Following this introduction, Section 2 provides a brief overview of relevant literature to provide a frame of reference for the study. Section 3 provides the research method used in the study, combining inductive and deductive approaches. Section 4 then provides the results of the study. Section 5 discusses some of the emergent themes and implications of this research. Finally, Section 6 concludes the research, presenting avenues for potential future research and highlighting major contributions and potential benefits of the study.

2. Literature Review

Information published on social media has had a major impact on trading and the stock market, and behavioral finance has demonstrated that emotions play an important role in financial decision-making. Extensive studies on financial news and social media attitudes have been conducted in the past, focusing on the association between Twitter sentiments and stock market behavior.

The integration of natural language processing (NLP) has gained popularity in stock market prediction [14–17]. Recent research has expanded our understanding of sentiment analysis across multiple applications [18–25]. Joshi and Tekchandani [26] examined sentiment prediction using 17,000 tweets employing multiple machine learning algorithms such as support vector machines (SVMs), maximum entropy, and Naive Bayes (NB). The N-gram sequence of words with unigrams, bigram, and a hybrid (unigram with bigram) feature was used, and SVM with the hybrid feature outperformed the other machine learning methods; however, there were concerns about the practicality of the linguistic feature, and the study highlights a limitation in the effectiveness of the linguistic feature in the sentiment analysis [27].

Bollen et al. [12] conducted a pioneering analysis of the correlation between collective mood states and the Dow Jones Industrial Average (DJIA). Two mood monitoring tools, Opinion Finder and Google Profile of Mood States, were used to generate six mood variations (calm, alert, sure, vital, kind, and happy). A Granger causality analysis and a Self-organizing Fuzzy Neural Network were used to analyze the public mood state discovered by employing mood-tracking methods to forecast changes in DJIA closing values. The experiment revealed that the DJIA was strongly connected to popular sentiments. Qian and Rasheed [28] proposed a DJIA index prediction model. The study used Hurst exponent to choose a time with the most predictable results and used Auto-Mutual Information and False Nearest Neighbor methods to select parameters which identify the training patterns. They conducted experiments using other machine learning models as well, such as Artificial Neural Networks, DT, K-Nearest Neighbor, and Ensemble methods. The Stacking Ensemble and Simple Voting methods were poorly performed; however, the study suggested that an ensemble of multiple classifiers will be useful for stock market prediction.

Yuan [25] explored the methods of sentiment classification using Twitter data, and three different types of methods of classification of sentiment analysis were discussed which were lexicon based, rule based, and machine learning based. The lexicon sentiments were discussed by applying Feature Scoring and Simple Word Count approach and concluded that the VADER Sentiment Lexicons and Bing Liu's Lexicon have been proven effective in Twitter sentiment analysis. Several questions have been raised regarding whether the time and effort used in applying the linguistic feature are worth it, as the improvement in the sentiment analysis results is insignificant, and whether using accuracy and precision is acceptable in evaluating sentiment analysis classification. This remains to be yet addressed and is seen as a limitation in the study. Urolagin [22] study on social media opinions and stock prices using Naïve Bayes and SVM classifiers identified that there is a relationship between the features of the tweets and the number of positive, neutral, and negative tweets. The study concludes that SVM performs better than the NB classifier on a tenfold cross-validation prediction. However, the major drawback of this study was conflicts between neutral and positive classes in both models, that the neutral and positive classes have high conflicts in both models. A similar limitation was observed in Mehta's and Pandya [29] lexicon-based opinion-mining approach emphasizing the need for addressing the class conflict. The study concluded that the SVM, NB, and the Neural Network approaches have achieved high accuracies and have the potential to be used in various applications.

Qasem et al. [11] study used Logistic Regression and Neural Network algorithms to compare sentiments related to technology stocks. Both the classifiers had the same overall accuracy of 58%, with significant conflicts between neutral and positive classes. This conflict might have arisen as a result of the automatic

selection of classes, particularly the neutral class, suggesting clustering methods that could be used to enhance model performance [30].

Ridhawi and Osman [31] study introduced an ensemble-based model that utilized multilayer perceptron, long short-term memory, and convolutional neural network models. This innovative approach achieved a remarkable next-hour prediction performance of 74.3%, emphasizing the synergy of sentiment analysis and financial data. Further, Mokhtari et al. [32] study explored the correlation between tweets and stock symbol trends, using long short-term memory, Bernoulli NB, and RF algorithms. The study found a significant relationship between Twitter sentiment and stock prices, emphasizing the correlation between stock market behavior and social media. Deveikyte et al. [33] study delved into the relationship between sentiment from financial news and tweets and FTSE100 movements. The study evidenced a correlation between sentiment and stock market movements, with sentiment from news headlines predicting market returns and negative sentiment from tweets correlating with lower volatility. Zaman et al. [34] research focused on incorporating sentiments related to external factors and historical data; the model achieved an accuracy of 87.2%, highlighting the role of social media in affecting IBM stocks. Multiple gaps were identified in the existing literature. Particularly, Alsing and Bahceci [35] used supervised and unsupervised machine learning algorithms to predict the stock market based on Twitter sentiments employing a limited number of tweets due to the restriction in retrieving Twitter data. The omission of emoticons, emojis, and slang in various studies [2, 36, 37] when carrying out sentiment analysis was recognized as a drawback limiting in capturing the full content of the tweet. Furthermore, there is a significant gap in the existing literature concerning the model evaluation. A predominant reliance on confusion matrix and accuracy score which is more suited to classification problems rather than assessing the performance of regression models based on prediction errors [38, 39], emphasizing the necessity for a more nuanced approach to model evaluation in stock market prediction research.

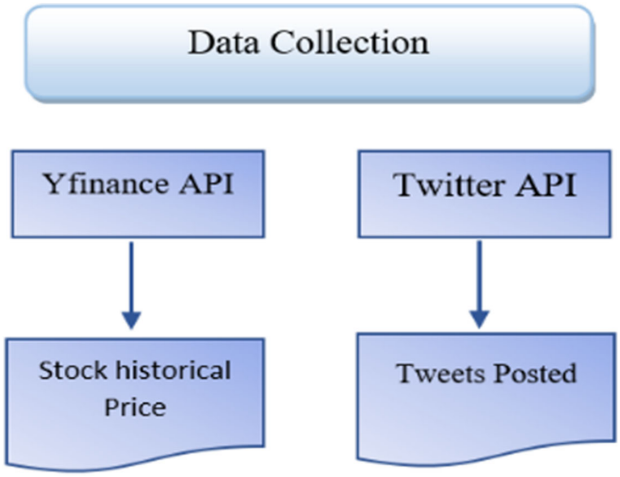
3. Research Method

This study uses a mixed-methods research approach to predict stock market prices using Twitter sentiment analysis. It was driven by a desire to gain an in-depth understanding of the intricate relationship between sentiments on social media and financial market behavior. By using a pragmatic-research mindset, it is found that the stock market dynamics are impacted by both quantitative and qualitative data and sentiment-based factors. As a result, using a mixed-methods approach is consistent with the research philosophy and enables a combined use of the benefit of quantitative and qualitative data sources. Both deductive and inductive research approaches are used in this study. Existing economic theories proposed a correlation between Twitter sentiments and the price of stocks, and so a deductive technique was first employed to create a hypothesis based on existing research and theory, arguing that Twitter sentiment had an impact on stock prices. Concurrently, an inductive method was applied to investigate unforeseeable factors and attitudes that may potentially impact market dynamics with no assumptions established from the beginning. Overall, deductive reasoning combined with inductive research was employed in this study to facilitate a comprehensive evaluation of the correlation between Twitter attitudes and stock market behavior.

3.1. Data collection

Two distinct datasets were employed to facilitate sentiment analysis and subsequently predict stock prices based on sentiment analysis outcomes. The study focuses on Tesla Inc., a multinational corporation traded on the New York Stock Exchange with a significant market capitalization of $577.43 billion. Tesla possesses a highly active Twitter presence, boasting over 18.2 million followers. Empirical evidence, as reported by various news sources [40], substantiates the impact of tweets related to Tesla and those posted by Elon Musk, the CEO of Tesla, on the company's stock market prices. This research aims to predict daily stock prices by leveraging sentiments expressed on the social media platform Twitter. To achieve this, tweets linked to Tesla and Elon Musk were systematically gathered from Twitter, while historical stock price data were sourced from the yFinance platform, as shown in Figure 1.

Figure 1
A framework for data collection



Twitter is one of the world's leading microblogging platforms, gaining global appeal for its ability to provide users with an environment to genuinely express their feelings, ideas, and address pertinent concerns [41]. In the context of this research, Twitter data were directly retrieved through a Python library known as Snscrape, which permits unhindered access to Twitter's application programming interface (API). Specifically, the data extraction encompassed critical attributes, including the number of likes, retweets, and tweet dates, in conjunction with the tweet content. The tweets containing the "#Tesla" hashtag were extracted within the time frame spanning from December 30, 2021, to November 2, 2022. Concurrently, tweets posted by Elon Musk were collected from December 30, 2020, to November 2, 2022. This data collection process yielded a total of 465,721 tweets associated with the "Tesla" hashtag and 6,251 tweets sourced from Elon Musk's Twitter profile.

The historical stock price data were retrieved using yFinance, an open-source tool that facilitates access to market data through Yahoo's publicly accessible API, as shown in Figure 2. The attributes closing price and the moving averages for 7, 20, and 50 days were retrieved for the dates from February 7, 2020, to November 7, 2022.

Figure 2
Historical stock data

	open	high	low	close	adjclose	volume	ticker	ma_7	ma_20	ma_50
2020-08-31	148.203339	166.713333	146.703339	166.106674	166.106674	355123200	TSLA	NaN	NaN	NaN
2020-09-01	167.380005	167.496674	156.836670	158.350006	158.350006	269523300	TSLA	NaN	NaN	NaN
2020-09-02	159.663330	159.679993	135.039993	149.123337	149.123337	288528300	TSLA	NaN	NaN	NaN
2020-09-03	135.743332	143.933334	134.000000	135.666672	135.666672	262788300	TSLA	NaN	NaN	NaN
2020-09-04	134.270004	142.666672	124.006668	139.440002	139.440002	330965700	TSLA	NaN	NaN	NaN
2020-09-08	118.666664	122.913330	109.959999	110.070000	110.070000	346397100	TSLA	NaN	NaN	NaN
2020-09-09	118.866669	123.000000	113.836670	122.093330	122.093330	238397400	TSLA	140.121432	NaN	NaN
2020-09-10	128.736664	132.996674	120.186668	123.779999	123.779999	254791800	TSLA	134.074764	NaN	NaN
2020-09-11	127.313332	127.500000	120.166664	124.239998	124.239998	182152500	TSLA	129.201905	NaN	NaN
2020-09-14	126.983330	140.000000	124.433334	139.873337	139.873337	249061800	TSLA	127.880477	NaN	NaN

3.2. Data preprocessing

Twitter data contains user-generated tweets that include unreadable emojis, unknown words, and characters. To perform sentiment analysis on the data, it is vital that the data are cleaned, and all irregularities are eliminated so that the original context and sentiments are captured [42]. Cleaning up the data and reducing noise in the dataset will help increase the performance of the sentiment analyzer and speed up the overall process [43]. In this study, data preprocessing was conducted to enhance the tweet dataset's quality for sentiment analysis. "@username" mentions, URLs, and special characters were removed from the tweets to ensure data cleanliness and model suitability. Additionally, spam tweets were identified and removed using a spam-check function, and duplicate tweets were managed by retaining the most engaged one. "Stop words" were also eliminated to improve sentiment analysis accuracy. Furthermore, it was observed that the dataset contained tweets in various languages, potentially complicating the use of the VADER sentiment analysis tool, primarily designed for English text. LangId, a language identification tool, was employed to filter out any non-English tweets to reduce linguistic inaccuracies.

3.3. Sentiment analyzer

The Natural Language Toolkit (NLTK), an open-source Python module, was used for sentiment analysis of the Twitter dataset. Developed for research in NLP, AI, cognitive science, and machine learning [44], NLTK encompasses various text processing capabilities, including sentiment analysis, which we utilized to evaluate the Twitter dataset. Within NLTK, a crucial component is the VADER tool. VADER relies on a lexicon containing over 7,500 words, each linked to sentiment scores, and extends its analysis to encompass Western-style emoticons, slang terms (e.g., "nah" and "yah"), and acronyms (e.g., "lol" and "rof"). Notably, VADER employs a Wisdom of the Crowd approach, gathering sentiment ratings from a group of individuals through Amazon's Mechanical Turk, a widely used crowdsourcing platform. This method enhances the tool's accuracy and reliability in assessing social media content, as demonstrated in previous studies [45, 46]. VADER takes into account various elements that can significantly impact sentiment in social media data, including emoticons, capitalization, slang usage, and text formatting. The VADER tool within the NLTK package was used in this research; it categorizes tweets as positive, negative, neutral, or compound, providing an overall sentiment score. The sentiment scores are based on an intensity scale ranging from −4 (severe negativity) to +4 (extreme positive). This comprehensive approach allowed for a deeper understanding of sentiment patterns within the Twitter dataset and their potential implications for the research objectives. Table 1 shows sentiment scores for sample words using VADER.

Table 1
Examples of the valence score from VADER

Words	Valence score	Words
Good	0.9	Good
Okay	1.9	Okay
Great	3.1	Great

The sentiment of the whole tweet can be calculated by adding up the valence score of each word in the tweet. VADER sums up the entire scores of features within the tweet and normalizes the final score to (−1, 1) using the normalizing function. The alpha is set to 15 to get the maximum expected value of x.

The standard threshold to classify a tweet as positive, negative, and neutral based on the compound score is as follows:

- Positive sentiment: compound score ≥ 0.05
- Neutral sentiment: compound score > -0.05 and < 0.05
- Negative sentiment: compound score ≤ -0.05

Two primary datasets were extracted: Twitter data and historical stock price data. The dataset consisted of the date, tweets, favorite count, retweet count, adjusted close values, and moving average features. To predict stock market prices based on the tweets' sentiments, this research employed six distinct machine learning algorithms. These algorithms were tested on two different sets of data. The final dataset was divided into two components before applying the machine learning models: a training set with 90% of the data and a testing set with 10% of the data. It is important to note that the

quality of the training and testing methods has a significant impact on the effectiveness of machine learning models [47]. A 50–50 split of training and testing data is used when dealing with closely related datasets. If the model fails to satisfy expectations, increasing the percentage of training data becomes vital because having < 50% training data may have an adverse effect on the model's testing outcomes [47]. Furthermore, data scaling was undertaken to transform the dataset into a Gaussian (normal) distribution, thereby optimizing the performance of the machine learning models. Scikit Learn's MinMaxScaler was employed to scale the final dataset, normalizing the data within the range of 0 to 1. A diverse set of six machine learning algorithms was utilized to predict stock prices based on Twitter sentiments. These algorithms encompass LR, DT, RF, XG Boost, SVR, and Facebook Prophet (FBS), providing a comprehensive approach to the prediction task.

4. Results

Two distinct Twitter datasets were examined. The original tranche of data, known as the "#Tesla Twitter raw dataset," had a sizable 465,721 tweets. It was reduced to a more manageable 9,402 tweets, though, following thorough data processing. Similar data-cleaning techniques were used for Elon Musk's Twitter dataset, which was reduced from an initial 6,251 tweets to 4,663 pertinent tweets. The VADER sentiment analysis tool was then used to perform sentiment analysis on these datasets. The VADER sentiment analysis tool generated valence scores for particular words within the tweets, indicating their level of positivity or negativity. These word-level ratings were combined and normalized to provide a composite sentiment score or compound score. Consider the following example tweet: "awesome tsla tesla," which received a compound score of 0.6249, a positive score of 0.672, a neutral score of 0.328, and a negative score of 0.0. The tool gives a neutral score when a word is not found in the VADER dictionary. The proportion of text that falls into each category of sentiment is represented by the positive, neutral, and negative scores, which add up to 1. The VADER sentiment analyzer tool was rigorously evaluated using diverse sentence variations, emoticons, and slang to assess its performance across different emotional contexts. The results showed promising outcomes, with sentences featuring exclamation marks and capitalized words attaining higher positive scores, while slang words like "sux" generated elevated negative scores. Table 2 shows the different variations to which the sentiment analyzer tool was applied.

4.1. The stock prediction results

The sentiments generated from tweets, along with historical closing prices, were employed to train and evaluate machine learning models. In the study, six distinct machine algorithms were used: LR, DT, RF Regressor, XG Boost, SVR, and FBS. The outcomes were evaluated to determine which machine algorithm produced the best results. Using these machine learning models, four distinct experiments were conducted in this research. The algorithms were first applied to two different Twitter datasets (#Tesla and Elon Musk's tweets) containing tweet sentiment and stock prices (adjusted closing price) to predict closing prices for seven days. Later, to improve accuracy, the moving average and adjusted closing prices were included.

The results, as shown in Figure 3, indicate that all the algorithms tested performed satisfactorily, with RF, SVR, and XG Boost outperforming the other three. On the contrary, the results displayed in Figures 4 and 5 indicate uncertainty about the model's effectiveness due to its deviation from expected outcomes. However, in this experiment, Prophet outperformed the other models, achieving better outcomes. Notably, the models made predictions considerably higher than the actual prices when the moving averages were not considered. Figure 6 shows that all algorithms work consistently and rather well, with XG Boost, RF, and SVR appearing as the top-performing approaches among the six algorithms tested. The use of moving averages, notably those covering 7, 20, and 50 days, is shown to significantly enhance prediction accuracy when compared to models without such moving averages.

5. Discussion

This study used two Twitter datasets: one sourced from Elon Musk's Twitter profile, which included his tweets and responses, and the other from tweets featuring the #Tesla hashtag. Tesla's CEO, Elon Musk, has a significant and prominent presence on social media, particularly Twitter. The changing social media landscape has given rise to "social media influencers," persons who shape user views, emotions, and actions through digital interactions [48]. This study focuses exclusively on Elon Musk's tweets, acknowledging his significant influence in promoting Tesla through Twitter conversations.

Furthermore, Elon Musk's tweets have had a significant impact on financial markets and individual investors' reactions and decisions. Musk's tweet acknowledging the encrypted messaging

Table 2
Sentiment scores for sample sentences and emotions

	Sentences	Negative	Neutral	Positive	Compound
1	The car is super cool	0.000	0.326	0.674	0.735
2	😊	0.000	0.522	0.478	0.671
3	☹	0.706	0.294	0.000	−0.340
4	The car is super cool!!!	0.000	0.298	0.702	0.795
5	The car is super cool!	0.000	0.316	0.684	0.757
6	The car is super COOL!	0.000	0.293	0.707	0.803
7	Tesla is extremely good	0.000	0.556	0.492	0.493
8	Tesla is moderately good	0.000	0.580	0.420	0.440
9	Tesla is good	0.000	0.508	0.444	0.440
10	Car is extremely good, but their service is horrible	0.343	0.506	0.151	−0.565
11	Today SUX!	0.779	0.221	0.000	−0.546
12	Today only kinda sux! But I'll get by lol	0.127	0.556	0.317	0.525

Figure 3
The outcomes of the #tesla tweets dataset with moving averages for 7, 20, and 50 days

	Date	Actual	RF	XGB	LR	FBS	SVR	DT
0	2022-10-26	224.639999	293.458796	315.367573	291.435972	252.915402	333.533908	280.899994
1	2022-10-27	225.089996	287.516667	294.162073	277.210626	248.652384	336.577944	289.913330
2	2022-10-28	228.520004	276.581268	282.815685	282.888510	251.219435	281.163519	216.759995
3	2022-10-29	228.520004	301.355764	281.745214	288.688838	252.193471	332.196893	335.016663
4	2022-10-30	228.520004	298.027101	312.564319	284.065353	252.473383	326.615559	352.260010
5	2022-10-31	227.539993	282.455699	289.234633	280.132949	252.517182	278.118342	232.229996
6	2022-11-01	227.820007	279.195633	283.065955	275.789400	244.187118	287.804428	232.229996

Figure 4
The outcomes of #tesla tweets dataset without moving averages

	Date	Actual	RF	XGB	LR	FBS	SVR	DT
0	2022-10-26	224.639999	218.934232	222.221929	218.008711	244.815165	221.187218	214.740005
1	2022-10-27	225.089996	220.225666	226.592591	220.419963	244.126867	220.389091	206.236664
2	2022-10-28	228.520004	222.023431	225.856841	220.605128	240.417103	222.382140	214.740005
3	2022-10-29	228.520004	216.224767	221.943977	221.394928	242.305885	222.340220	220.583328
4	2022-10-30	228.520004	225.459964	228.590940	221.308017	248.377639	223.620550	223.656662
5	2022-10-31	227.539993	227.099800	229.104373	225.366530	245.559694	224.435056	237.036667
6	2022-11-01	227.820007	226.528800	228.583576	227.649674	245.268221	226.903819	236.580002

Figure 5
The outcomes of Elon Musk's tweets dataset without moving averages for 7, 20 and 50 days

	Date	Actual	RF	XGB	LR	FBS	SVR	DT
0	2022-10-26	224.639999	265.351934	265.049508	268.754588	244.815165	262.154765	218.429993
1	2022-10-27	225.089996	267.491098	257.880792	278.497532	244.126867	275.065362	220.190002
2	2022-10-28	228.520004	282.207467	257.163460	270.749769	240.417103	262.631312	225.673340
3	2022-10-29	228.520004	260.551100	286.823274	261.684025	242.305885	256.896867	246.066666
4	2022-10-30	228.520004	253.667703	286.303636	274.038423	248.377639	265.537724	210.283340
5	2022-10-31	227.539993	295.621929	298.783568	273.604770	245.559694	266.822847	281.666656
6	2022-11-01	227.820007	296.643099	317.123045	277.888353	245.268221	268.879990	335.016663

Figure 6
The outcomes of Elon Musk's tweets dataset without moving averages

	Date	Actual	RF	XGB	LR	FBS	SVR	DT
0	2022-10-26	224.639999	223.837934	234.724005	219.218071	226.646107	234.863176	237.036667
1	2022-10-27	225.089996	225.582434	228.393421	224.382094	217.437145	227.001362	237.036667
2	2022-10-28	228.520004	225.206034	225.482905	223.340516	220.860317	223.109199	237.036667
3	2022-10-29	228.520004	223.986400	223.875152	222.427295	221.243254	227.378216	222.419998
4	2022-10-30	228.520004	225.966234	229.489305	223.602151	218.480177	225.723886	237.036667
5	2022-10-31	227.539993	225.014467	225.760771	226.983358	219.670650	225.968558	237.036667
6	2022-11-01	227.820007	225.718167	236.726774	231.577873	210.831893	226.423731	237.036667

Figure 7
The distribution of Elon Musk's tweets between positive, negative, and neural sentiment score

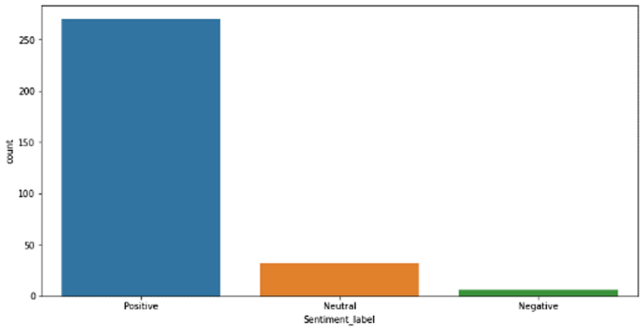
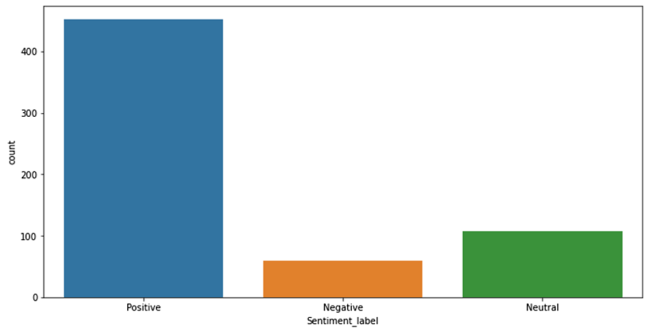


Figure 8
The distribution of #tesla tweets between positive, negative, and neural sentiment scores



service Signal exhibits this, as it significantly increased investor interest in Signal Advance shares. This rise in demand caused Signal Advance's market value to soar from $55 million to more than $3 billion [49]. This noteworthy incident demonstrates the impact of Elon Musk's tweets on financial markets and individual investors' choices [50].

Musk's Twitter activity generated a dataset with a variety of attitudes, including neutral, positive, and negative emotions as depicted in Figure 7; it comprised a total of 453 positive, 107 neutral, and 59 negative tweets. In contrast, the dataset associated with tweets containing the #Tesla hashtag, as seen in Figure 8, demonstrated a count of 270 positive, 32 neutral, and 6 negative tweets. The variation in sentiment distribution can be traced to the aggregate of tweets for each date, an essential requirement for the functioning of the stock prediction model. Consequently, the aggregated sentiment scores for most dates in the Tesla dataset resulted in a predominance of positive tweets and comparatively lower count of negative tweets. Notably, when words in tweets do not match words in the VADER dictionary, neutral values are automatically assigned, potentially leading to inaccuracies in sentiment scores and contributing to an elevation in positive tweets within the dataset. It is important to acknowledge VADER's limitations, as identified through performance evaluation, particularly when dealing with complex tweets and words that are not in its vocabulary, which results in incorrect sentiment scores. However, despite these challenges, VADER produced satisfactory results for the majority of tweets, aligning with the outcomes of previous researchers who have applied this tool to analyze social media data [45,46].

5.1. Stock prediction analysis

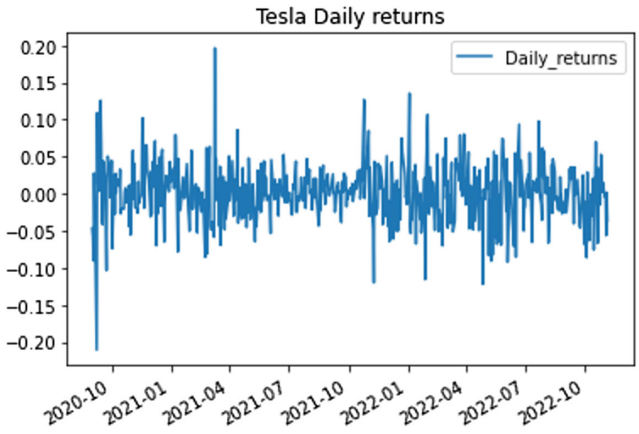
The historical stock data (adjusted close price) was retrieved using the yFinance tool from Yahoo Finance. The adjusted close is the closing price of the stock after adjusting for any corporate action events such as dividends and stock splits. Figure 9 shows the adjusted closing of Tesla from October 1, 2020, to November 1, 2022. The prices have risen steadily from 2020 to 2022, and from January 1, 2022, a steady drop in the prices has been seen. The peak was during the pandemic when the Tesla shares reached the highest ever. In this study, data from July 1, 2022, to November 1, 2022, are considered.

Figure 9
The adjusted closing price from January 10, 2020, to January 11, 2022

As shown in Figure 9, a significant decline in prices is observed from September 1, 2022, to October 15, 2022. This decline can be attributed to investors divesting their shares following Tesla's announcement of a $5 billion stock offering and a substantial shareholder liquidating their holdings [51]. This underscores the impact of investor reactions to information on the stock market. In situations where a greater number of investors opt to sell stocks compared to those buying, it results in a notable decline in stock prices.

Figure 10 shows the daily returns of Tesla between October 1, 2020, and November 1, 2022, indicating a notable fluctuation in daily returns. The fluctuation is consistent with the movements observed in the daily adjusted stock prices. Following the generation of sentiment scores, a comparative analysis with stock prices was performed to determine whether the two variables had a visible positive or negative correlation. A correlation coefficient was calculated between the two variables. The correlation coefficient measures the relationship between two variables by comparing the degree of change in one variable to the degree of change in the other variable, either in the same direction (positive correlation) or in the opposite direction (negative correlation). Furthermore, correlation can be used to determine the strength of the relationship [52]. #Tesla tweets have a correlation of 8.6%, and Elon Musk's tweets have a correlation of −11.35%, which suggests that #tesla datasets are positively correlated, and Musk's

Figure 10
The daily returns of Tesla from January 10, 2020, to January 11, 2022



tweets data are negatively correlated. The correlation coefficient results suggest that the adjusted closing price and the sentiment score have a weak linear relationship. Figures 11 and 12 show the movement of the compound (overall sentiment score) and the adjusted closing price for the #tesla dataset during the period. It suggests that the adjusted closing prices steadily increase and steadily fall; however, the sentiment scores fluctuate heavily. Deveikyte et al. [33] analysis coincides with Musk's Twitter dataset in this study, highlighting a significant negative correlation between positive tweets and the subsequent day's market volatility. The leverage effect can be observed that during periods of declining stock prices, the impact on the volatility of stock prices is more pronounced than during periods of rising prices [53].

#Tesla tweets dataset is 22.01% positively correlated with the moving averages calculated for 7 days and 14.74% positively correlated with moving averages calculated for 20 days. Elon Musk's tweets are −12.18% negatively correlated with moving averages calculated for 7 days and −13.79% negatively correlated

Figure 11
The movement of adjusted closing price of Tesla and the compound score between the period January 07, 2022, and January 11, 2022 (Tesla tweets)

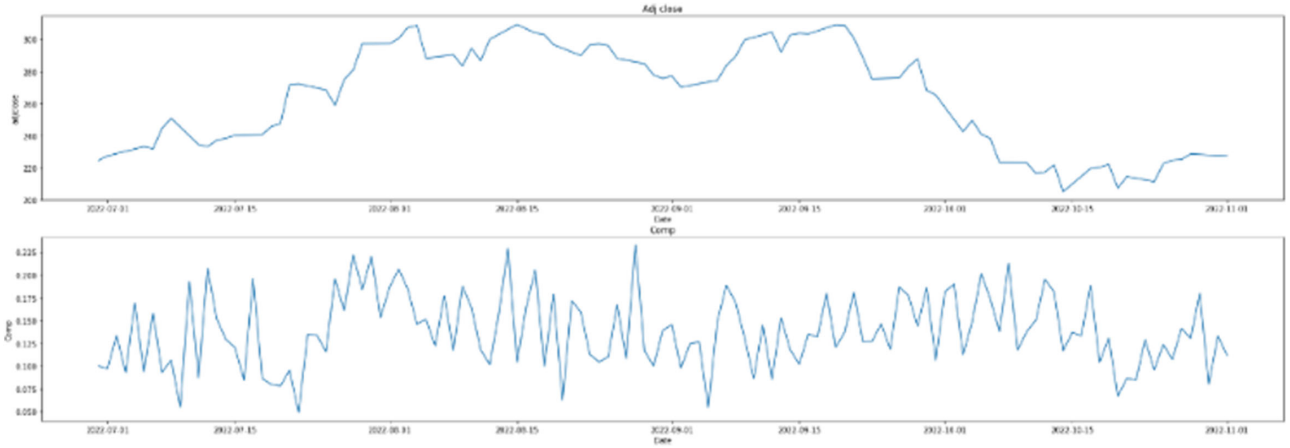
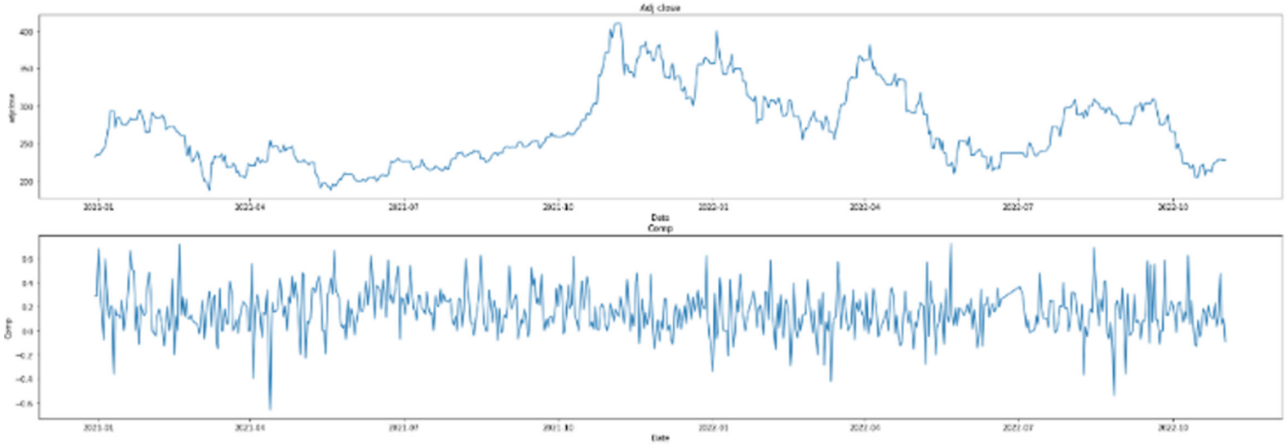


Figure 12
The movement of adjusted closing price of Tesla and the compound score between the period January 1, 2020, and January 11, 2022 (Elon Musk's tweets)



with moving averages calculated for 20 days. The correlation coefficient reduces with the increase in the number of days.

The study employed various machine learning algorithms to predict stock prices using sentiment analysis derived from Twitter data. Additionally, moving averages were integrated to enhance the accuracy of price predictions. The six algorithms used in this research included LR, DT, RF Regressor, SVR, XG Boost, and FBS.

Model evaluation was conducted using three key metrics: mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). In the case of the #Tesla Twitter dataset, RF exhibited the lowest MAE, achieving a score of 2.137, followed by SVR with an MAE of 3.493. Interestingly, Prophet yielded the highest MAE score at 23.358. Conversely, in Elon Musk's Twitter dataset, XG Boost attained the highest MAE score of 2.223, while Prophet achieved the lowest score at 17.174. Without the incorporation of moving averages in both datasets, Prophet consistently obtained the highest MAE scores, suggesting a greater deviation of its predictions from actual values. Asgarov [54] research employed the MAE metric to evaluate the performance of the machine learning model. Similar datasets were used to predict the stock market, and the obtained MAE values were 9.93 and 2.47. These results indicate a consistent performance, demonstrating the model's capability to make reasonable predictions.

MSE, as a measure of the squared distance between actual and predicted values, was also employed for evaluation. RF and XG Boost emerged with the lowest MSE scores for both the #Tesla and Elon Musk's datasets, indicating that data points in these datasets were closely distributed around the mean. Additionally, these models displayed fewer errors and exhibited less skewed predictions compared to other algorithms. Prophet consistently outperformed other models in terms of MSE for datasets without moving averages. Interestingly, SVR, which performed well in MAE, was outperformed by LR in the MSE score, implying the possible presence of a significant error in SVR predictions. Furthermore, RF and XG Boost achieved the lowest RMSE in datasets with moving averages, suggesting that RF and XG Boost best fit the model and achieved the highest accuracy. Conversely, Prophet demonstrated better performance than other algorithms in datasets without moving averages; however, its predictions consistently exceeded actual prices.

The study's findings revealed that classification models, particularly RF, SVR, XG Boost, and LR, effectively predicted the direction of stock price movements, albeit with variances compared to actual results. Notably, LR closely competed with RF, SVR, and XG Boost in terms of predictive performance. While previous research favored SVR as an optimal choice for stock price prediction [55], this study suggests that RF outperformed SVR, potentially influenced by dataset characteristics. Given the highly volatile nature of Tesla's stock prices, which experienced fluctuations from $362 on December 29, 2021, to $204 on October 15, 2022, before rebounding to $227 on November 1, 2022, the algorithms' performance may have been impacted by these market dynamics.

6. Conclusion

In conclusion, this research highlights the potential value of integrating social media data as a feature in machine learning models for stock market prediction. However, a critical consideration arises regarding the reliability of the data, with a substantial portion of tweets identified as spam, unrelated, or unusable. The geographical origin of tweets further complicates the data's validity, as tweets from irrelevant locations may not significantly impact stock market prices. Despite these challenges, various machine learning models were rigorously tested using four datasets, revealing that the high volatility of Tesla stock prices posed a challenge for all models. Among the tested models, XG Boost and RF Regressor emerged as the most effective, particularly when the datasets included moving averages. The incorporation of Elon Musk's tweets demonstrated an immediate short-term improvement in the model's performance, albeit limited by the influence of other external market factors. Notably, combining Musk's tweets with moving averages yielded enhanced results. The research suggests that further refinements could be achieved by introducing additional financial metrics such as beta, price-to-earnings (P/E) ratio, dividend yield, and profit margin into the predictive model.

While correlation coefficients indicated a positive relationship between #Tesla tweets and stock prices and a negative correlation with Elon Musk's tweets, these correlations were small, ranging below 15%. The study acknowledges the numerous internal and external factors influencing stock prices, encompassing financial indicators, economic conditions, political stability, and technological advancements. Consequently, social media sentiments have an impact but cannot be solely depended on for accurate stock price predictions.

In conclusion, this research highlights the potential of social media data to serve as an additional indicator in combination with traditional financial indicators, improving the overall model and contributing to a more holistic understanding of stock market dynamics. The RF Regressor and XG Boost models, particularly when moving averages were included, showed promise for future investigation. Future research could look into the geographical dimension of tweets and expand the analysis to include various companies, improving the resilience and usefulness of predictive models in the ever-changing landscape of stock market prediction.

While this research contributes valuable insights into the relationship between social media sentiments and stock market dynamics, several limitations must be acknowledged. Firstly, the reliability of social media data, particularly from Twitter, poses a significant challenge. The prevalence of spam, unrelated content, and unusable data within the dataset raises concerns about its accuracy and representation. To address this limitation, future research should prioritize the use of advanced filtering mechanisms to enhance the reliability of social media data.

Another notable limitation pertains to the geographical dimension of tweets. The study identified tweets originating from different countries, potentially impacting the relevance of the data to the specific stock market in focus. Strategies should be developed to incorporate this geographical dimension, ensuring regional influences in the social media data are considered.

The high volatility of stock prices emerged as a substantial challenge, affecting the performance of machine learning models. To overcome this limitation, future research should explore using additional features to reduce the effect of volatility and enhance model resilience in the face of rapid price fluctuations.

Furthermore, the study revealed moderate correlation coefficients between social media sentiments and stock prices. While positive correlations were observed with #Tesla tweets and negative correlations with Elon Musk's tweets, the nature of these correlations suggests limitations in the predictive power of social media sentiments alone. The immediate short-term impact of Elon Musk's tweets on model performance also emerged as a limitation. While the inclusion of Musk's tweets improved results, it was effective only in the immediate short run. This underscores

the constraints of social media data in predicting stock price movements and emphasizes the significance of external factors beyond sentiment analysis.

Additionally, the study acknowledged the potential for improvement by incorporating additional financial metrics, such as beta, P/E ratio, dividend yield, and profit margin. This limitation highlights the need for a more comprehensive approach that integrates both social media insights and traditional financial indicators for a more accurate and holistic approach.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

References

[1] Wong, B. (2023). *Top social media statistics and trends of 2023.* Retrieved from: https://www.forbes.com/advisor/business/social-media-statistics/

[2] Mehta, P., Pandya, S., & Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, *7*, e476. https://doi.org/10.7717/peerj-cs.476

[3] Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics*, *10*(21), 2717. https://doi.org/10.3390/electronics10212717

[4] Wang, Z., Ho, S. B., & Lin, Z. (2018). Stock market prediction analysis by incorporating social and news opinion and sentiment. In *2018 IEEE International Conference on Data Mining Workshops*, 1375–1380. https://doi.org/10.1109/ICDMW.2018.00195

[5] Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 3433–3456. https://doi.org/10.1007/s12652-020-01839-w

[6] Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, *17*(1), 59–82. http://doi.org/10.1257/089533003321164958

[7] Schwartz, R. A. (1970). Efficient capital markets: A review of theory and empirical work: Discussion. *The Journal of Finance*, *25*(2), 421–423. https://doi.org/10.2307/2325488

[8] Degutis, A., & Novickytė, L. (2014). The efficient market hypothesis: A critical review of literature and methodology. *Ekonomika*, *93*(2), 7–23. https://doi.org/10.15388/Ekon.2014.2.3549

[9] Naseer, M., & Bin Tariq, D. Y. (2015). The efficient market hypothesis: A critical review of the literature. *The IUP Journal of Financial Risk Management*, *12*(4), 48–63. https://ssrn.com/abstract=2714844

[10] Omar, A. B., Huang, S., Salameh, A. A., Khurram, H., & Fareed, M. (2022). Stock market forecasting using the random forest and deep neural network models before and during the COVID-19 period. *Frontiers in Environmental Science*, *10*, 917047. https://doi.org/10.3389/fenvs.2022.917047

[11] Qasem, M., Thulasiram, R., & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics*, 834–840. https://doi.org/10.1109/ICACCI.2015.7275714

[12] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[13] Virtanen, I., & Yli-Olli, P. (1987). Forecasting stock market prices in a thin security market. *Omega*, *15*(2), 145–155. https://doi.org/10.1016/0305-0483(87)90029-6

[14] Bhardwaj, A., Narayan, Y., Vanraj, Pawan, & Dutta, M. (2015). Sentiment analysis for Indian stock market prediction using Sensex and Nifty. *Procedia Computer Science*, *70*, 85–91. https://doi.org/10.1016/j.procs.2015.10.043

[15] Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: Preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*, 1. https://doi.org/10.1186/s40649-020-00080-x

[16] Huang, J. Y., & Liu, J. H. (2020). Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting*, *39*(1), 104–116. https://doi.org/10.1002/for.2616

[17] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559. https://doi.org/10.1108/IntR-06-2012-0114

[18] Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

[19] Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the effects of news sentiments on the stock market. In *2018 IEEE International Conference on Big Data,* 4705–4708. https://doi.org/10.1109/BigData.2018.8621884

[20] Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data,* 77–88. https://doi.org/10.1007/978-3-642-39146-0_8

[21] Sul, H. K., Dennis, A. R., & Yuan, L. (2017). Trading on Twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, *48*(3), 454–488. https://doi.org/10.1111/deci.12229

[22] Urolagin, S. (2017). Text mining of Tweet for sentiment classification and association with stock prices. In *2017 International Conference on Computer and Applications,* 384–388. https://doi.org/10.1109/COMAPP.2017.8079788

[23] Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016). Stock market prediction using machine learning techniques. In *2016 3rd International Conference on Computer and Information Sciences,* 322–327. https://doi.org/10.1109/ICCOINS.2016.7783235

[24] Wang, Y., & Wang, Y. (2016). Using social media mining technology to assist in price prediction of stock market. In *2016 IEEE International Conference On Big Data Analysis,* 1–4. https://doi.org/10.1109/ICBDA.2016.7509794

[25] Yuan, B. (2016). *Sentiment analysis of Twitter data*. Doctoral Dissertation, Rensselaer Polytechnic Institute. Retrieved from: https://www.cs.rpi.edu/~szymansk/theses/bo.ms.16.pdf

[26] Joshi, R., & Tekchandani, R. (2016). Comparative analysis of Twitter data using supervised classifiers. *International Conference on Inventive Computation Technologies, 3*, 1–6. https://doi.org/10.1109/INVENTIVE.2016.7830089

[27] Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 695–704. https://doi.org/10.1145/1963405.1963503

[28] Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*, 25–33. https://doi.org/10.1007/s10489-006-0001-7

[29] Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific & Technology Research*, *9*(2), 601–609.

[30] Li, G., & Liu, F. (2010). A clustering-based approach on sentiment analysis. In *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, 331–337. https://doi.org/10.1109/ISKE.2010.5680859

[31] Ridhawi, M. A., & Osman, H. A. (2023). Stock market prediction from sentiment and financial stock data using machine learning. In *Proceedings of the 36th Canadian Conference on Artificial Intelligence.* https://doi.org/10.21428/594757db.40c1a462

[32] Mokhtari, M., Seraj, A., Saeedi, N., & Karshenas, A. (2023). The impact of Twitter sentiments on stock market trends. *arXiv Preprint: 2302.07244*. https://doi.org/10.48550/arXiv.2302.07244

[33] Deveikyte, J., Geman, H., Piccari, C., & Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, *5*, 836809. https://doi.org/10.3389/frai.2022.836809

[34] Zaman, N., Ghazanfar, M. A., Anwar, M., Lee, S. W., Qazi, N., Karimi, A., & Javed, A. (2023). Stock market prediction based on machine learning and social sentiment analysis. *TechRxiv Preprint*. https://doi.org/10.36227/techrxiv.22315069.v1

[35] Alsing, O., & Bahceci, O. (2015). *Stock market prediction using social media analysis.* Retrieved from: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-166448

[36] Mihir, A., Sumedh, N., & Anmol, K. (2020). Stock analysis using sentiment analysis and machine learning. *International Journal of Innovative Research in Technology*, *6*(11), 123–126.

[37] Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System,* 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659

[38] Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv Preprint: 1809.03006*. https://doi.org/10.48550/arXiv.1809.03006

[39] de Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J., & Moguerza, J. M. (2022). General performance score for classification problems. *Applied Intelligence*, *52*(10), 12049–12063. https://doi.org/10.1007/s10489-021-03041-7

[40] Reuters. (2021), *How tweets by Tesla's Elon Musk have moved markets.* Retrieved from: https://www.reuters.com/business/finance/how-tweets-by-teslas-elon-musk-have-moved-markets-2021-11-08/

[41] Spangler, E., & Smith, B. (2022). Let them tweet cake: Estimating public dissent using Twitter. *Defence and Peace Economics*, *33*(3), 327–346. https://doi.org/10.1080/10242694.2020.1865042

[42] Hemalatha, I., Varma, G. S., & Govardhan, A. (2012). Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science*, *1*(2), 58–61.

[43] Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

[44] Yao, J. (2019). Automated sentiment analysis of text data with NLTK. *Journal of Physics: Conference Series*, *1187*(5), 052020. https://doi.org/10.1088/1742-6596/1187/5/052020

[45] Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*, 1–6. http://doi.org/10.51983/ajcst-2019.8.S2.2037

[46] Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*.

[47] Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, *4*, 31. https://doi.org/10.1186/1755-8794-4-31

[48] Kim, J., & Kim, M. (2022). Rise of social media influencers as a new marketing channel: Focusing on the roles of psychological well-being and perceived social responsibility among consumers. *International Journal of Environmental Research and Public Health*, *19*(4), 2362. https://doi.org/10.3390/ijerph19042362

[49] DeCambre, M. (2021). *Why an Elon Musk tweet led to a 5,675% surge in Signal Advance's stock.* Retrieved from: https://www.marketwatch.com/story/why-an-elon-musk-tweet-led-to-a-5-675-surge-in-health-care-stock-signal-advance-11610400141

[50] Ante, L. (2023). How Elon Musk's Twitter activity moves cryptocurrency markets. *Technological Forecasting and Social Change*, *186*, 122112. https://doi.org/10.1016/j.techfore.2022.122112

[51] Kilgore, T. (2020). *Tesla stock suffers biggest-ever drop as it starts its second bear market this year.* Retrieved from: https://www.marketwatch.com/story/tesla-stock-tumbles-toward-2nd-bear-market-in-6-months-2020-09-08

[52] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

[53] Aït-Sahalia, Y., Fan, J., & Li, Y. (2013). The leverage effect puzzle: Disentangling sources of bias at high frequency. *Journal of Financial Economics*, *109*(1), 224–249. https://doi.org/10.1016/j.jfineco.2013.02.018

[54] Asgarov, A. (2023). Predicting financial market trends using time series analysis and natural language processing. *arXiv Preprint: 2309.00136*. https://doi.org/10.48550/arXiv.2309.00136

[55] Chen, S., Gao, T., He, Y., & Jin, Y. (2019). Predicting the stock price movement by social media analysis. *Journal of Data Analysis and Information Processing*, *7*(4), 295–305. https://doi.org/10.4236/jdaip.2019.74017

How to Cite: Kandasamy, T., & Bechkoum, K. (2024). Impact of Social Media Sentiments on Stock Market Behavior: A Machine Learning Approach to Analyzing Market Dynamics. *Journal of Comprehensive Business Administration Research*. https://doi.org/10.47852/bonviewJCBAR42022006

Appendix

Abbreviations

AI	Artificial intelligence
API	Application programing interface
BoW	Bag-of-words
CSV	Comma Separated Values
DJIA	Dow Jones Industrial Average
DT	Decision Tree
EMH	Efficient market hypothesis
FBS	Facebook Prophet
FEX	Foreign Exchange
GPOMS	Google-Profile of Mood States
HTML	Hypertext Markup Language
LR	Logistic Regression
MA	Moving Average
MACD	Moving Average Convergence/ Divergence
ML	Machine learning
MLP	Multilayer perceptron
MSE	Mean squared error
NASDAQ	National Association of Securities Dealers Automated Quotations
NB	Naïve Bayes
NLP	Natural language processing
NLTK	Natural Language Toolkit
Regex	Regular expressions
RF	Random Forest
RMSE	Root mean square error
SLP	Single layer Perceptron
SVM	Support vector machine
SVR	Support Vector Regressor
TF	Term frequency
TF-IDF	TF inverse document frequency
URL	Uniform Resource Locator
VADER	Valence Aware Dictionary and Sentiment Reasoner
WL Wavelet algorithms	Wavelet algorithms
WotC	Wisdom of the Crowd
XG Boost	Extreme Gradient Boosting