Understanding the Impact of Social Media Sentiment on Financial Decision-making within the Stock Market: A Deep Learning Computational Analysis

MSc in Science in AI for Business (MSCAIBUS1)

Osaigbovo Daniel Adoghe
Student ID: x23139013

School of Computing
National College of Ireland

Supervisor: Dr. Devanshu Anand

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Osaigbovo Daniel Adoghe
Student ID:	x23139013
Programme:	MSc in Science in AI for Business (MSCAIBUS1)
Year:	2024
Module:	MSc Research Project
Supervisor:	Devanshu Anand
Submission Due Date:	20/12/2018
Project Title:	Understanding the Impact of Social Media Sentiment on Financial Decision-making within the Stock Market: A Deep Learning Computational Analysis
Word Count:	6022
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL Internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use another author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Osaigbovo Daniel Adoghe
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on a computer.	□

Assignments that are submitted to the Programme Coordinator's office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Understanding the Impact of Social Media Sentiment on Financial Decision-making within the Stock Market: A Deep Learning Computational Analysis

Osaigbovo Daniel Adoghe
x23139013

Abstract

This study considers an innovative way to explore social media sentiment analysis for stock market prediction through the use of Generative Adversarial Networks (GANs) to improve the accuracy of the forecasting models. Traditional financial theories like the Efficient Market Hypothesis (EMH), and the Random Walk (RW) theory have often overlooked the psychological and behavioral aspects of market dynamics that drive stock prices. Thus, our study incorporates the psychological component through sentiment data expressed in X (formerly known as Twitter) by designing three predictive models; Long Short-Term Memory, Random Forest, and GAN. These models were subsequently evaluated against Tesla (TSLA) and Amazon (AMZN) stock data, focusing on some major performance metrics such as accuracy, precision, and recall. In this respect, the GAN model demonstrated superior performance with an accuracy of 82.67%, precision of 70.21%, and recall of 81.11% for TSLA, and accuracy of 84.21%, precision of 85.71%, and recall of 75.00% for AMZN. In comparison, the LSTM model achieved an accuracy of 66.67% for TSLA and 53.84% for AMZN, while the RF model achieved 56.86% for TSLA and 54.00% for AMZN. This research not only contributes to the evolution of computational finance but also accentuates how decisive behavioral economics can be in understanding and predicting market trends. The results subsequently indicate that incorporating social media sentiment increases substantially the predictive power of financial models, therefore offering a more nuanced approach toward market analysis.

Keywords: *Social Media Sentiment, Stock Market Prediction, Generative Adversarial Networks (GAN), Long Short-Term Memory(LSTM), Random Forest (RF), Behavioral Economics, Tesla (TSLA), Amazon (AMZN), Financial Forecasting, Computational Finance.*

1 Introduction

Stock market prediction has seemed a captivating field for both researchers and investors for the best part of the last two decades, owing to the large financial returns that could be accumulated from accurately predicted stock market movement. Traditional financial theories like the Efficient Market Hypothesis and the Random Walk theory have dominated for a very long time giving the impression that stock prices reflected all known information and so moved in a random path that could not reliably be predicted. However, these theories have been criticized for their limitations in explaining the non-linear and volatile nature of the market and for disregarding the psychological and behavioral factors that affect investment decisions by individual investors (Chopra and Sharma (2021) and Pang et al. (2020)).

The recent growth of social media platforms like X has added a new dimension to stock market analysis. The massive user-generated data on these platforms serve as a real-time data stream that can be analyzed to ascertain public opinion. Sentiment management includes subjective information extraction from textual data to establish the mood and opinion of the public on almost any topic, including how a stock is performing Ravi and Ravi (2015). Research has proven that the sentiment of social media users can significantly influence stock prices and hence provide predictive insights that otherwise could not be deducted by traditional financial models Rao and Srivastava (2012).

The integration of machine learning (ML) techniques especially advanced models, such as Generative Adversarial Networks (GANs) has tremendously impacted the landscape of stock market prediction. GANs are particularly known for creating synthetic data that mirrors the complexities of real-world data and subsequently enhances the robustness of the predictive models. Such models are excellent at capturing the intricate and nonlinear patterns present in financial data usually missed by the known traditional models Goodfellow et al. (2014)

Research Question: Can deep learning models accurately predict stock market movements by analyzing social media trends and sentiments from X (formerly known as Twitter)? This study aims to explore the innovative application of GAN with sentiment analysis results from social media particularly Twitter in predicting market dynamics and trends. In this study, 2 major companies, Tesla Inc. (TSLA) and Amazon Inc. (AMZN), have been selected because of their huge influence on the market and the large volume of sentiment data available. This investigation shall be guided by the notion that market sentiment, mostly reflected in social media, can increasingly provide valuable insights against those captured through conventional financial indicators.

The significance of the study lies in its attempt to help bridge the gap between traditional financial theories and modern computational techniques, which more accurately explain market dynamics influenced by public sentiment. The study contributes to this ever-growing field of financial analytics by demonstrating how the integration of social media sentiment analysis with cutting-edge machine learning techniques can significantly enhance the accuracy of stock market predictions.

Literature Review: survey previous work on different methods applied toward the prediction of stock markets; these mostly revolve around sentiment analysis and machine learning models.
Methodology: details of the research design, data collection methods, and preprocessing and also the development of ML models used, along with which evaluation metrics should be used to measure performance.
Design Specification: Outlines the architecture and technical specification of the models, including any new algorithms proposed at a high level of detail.
Implementation: describes the practical steps involved in the development of the models; tools and languages used are described.

Evaluation: present the results after the study is conducted with the use of statistical tools and aids for visual performance analysis of the models.
Conclusion and Future Work: The final section sums up the main findings of the study and assesses its success in answering the research questions. The discussion gives the implications of the research, outlines its limitations, and shows the directions for future work.

2 Related Work

The role that machine learning has taken in financial forecasting has completely revolutionized the field through advanced predictive capabilities that go beyond traditional statistical methods. Traditional approaches are most often limited by the linearity assumptions and inability to deal with large-scale, high-dimensional data and are now being supplemented or overridden by techniques of ML Chopra and Sharma (2021). Machine learning models perform optimally in revealing those complex, nonlinear relationships within big data, which turns out useful in the financial domain due to the highly volatile nature of market data and the myriad of factors influencing it.

2.1 Hybrid Deep Learning Models

Several studies have been carried out in recent times exploring the use of hybrid models in financial and stock market forecasting. In 2022, Huang et al. (2023) introduced a new approach using genetic algorithms combined with deep learning to predict the stock price accurately. The use of feature selection by GA helps the model identify the relevant variables and refine the predictive capabilities of the model. This hybrid approach shows the power of leveraging the unique strength of different machine learning techniques.

Chandola et al. (2023) also proposed a hybrid model that combined Word2Vec with Long Short-Term Memory neural networks. for numerical data on stock series and news headlines, to provide a comprehensive overview of different variables that could affect stock prices. Similarly, Huang et al. (2022) and Nijaguna (2023) developed hybrid models that combine various neural network architectures, such as LSTM, Support Vector Machine (SVM), and Empirical Mode Decomposition(EMD) to handle more complex and non-linear data capturing temporal patterns effectively

2.2 BERT and GANs for Stock Prediction

Sonkiya et al. (2021) integrated a fine-tuned BERT model into generative adversarial networks for the prediction of stock prices, focusing on Apple Inc. shares. This model utilized finBERT for sentiment analysis over financial texts and comprehensively understood the market sentiments and their impacts on stock prices. The novelty of this work lies in the integration of NLP and GANs and in demonstrating how qualitative data of sentiments can significantly improve traditional quantitative analysis.

2.3 Application of RNN, LSTM and Transfer Learning

Several studies have indicated that Recurrent Neural Networks (RNN) and its variants, such as LSTM are efficient in handling time series data for stock market prediction. Lakshmanarao et al. (2022) demonstrated the capabilities of Bi-directional LSTM, which increases the prediction accuracies by studying the stock data in both forward and backward directions. Sutradhar et al. (2021) drew attention to the fact that the capabilities of LSTM architecture can efficiently deal with long-term dependencies during the movement of stock prices.

Chen et al. (2022) applied the concept of transfer learning in stock price prediction. They used historical data of stock prices and corporate trading to train LSTM networks, outlining the effectiveness of transfer learning by fine-tuning pre-trained models on a totally new but related dataset. This strategy saves computation time while enhancing model performance, especially when data availability is limited. Similarly, Zhang et al. (2022) provided valuable insights into how LSTM could be used for stock market prediction by analyzing various influential market factors. While promising in its approach, the research also outlined some challenges of dealing with complex and noisy data, which have to be carefully managed to optimize the predictive capabilities of machine learning models.

2.4 Emotion/sentiment Analysis and Market Influence

Wang (2023)'s research focuses on understanding how statements from public figures, especially on social media, could move the stock market. The sentiment was quantified with a BERT-based model and correlated with changes in stock prices to understand how investor behavior may be swayed by influential public commentary. This puts more emphasis on the importance of sentiment analysis in financial forecasting.

Kumar et al. (2022) used a composite model by integrating Support Vector Machine (SVM), LSTM, and Gated Recurrent Unit (GRU) models to analyze the emotional sentiment expressed in news and social media texts with regard to correlating these sentiments with stock market behavior. This would provide an intricate understanding of how public emotion actually influences financial markets.

2.5 Integration of Diverse Data Sources and Social Media Sentiment Analysis

The potential of social media sentiment analysis concerning stock prediction has been in several studies. Mehta et al. (2021) demonstrated that real-time sentiment data from sources like Twitter, combined with traditional financial data, could greatly improve the predictive accuracy of stock models. Lim and Yeo (2020) also put forward a reason for the role of social media sentiment in stock market prediction by citing early indicators of market movements derived from public sentiment.

The integration of multiple data sources, including financial news, social media data, and technical indicators has been a reoccurring theme in recent research with Tiwari et al. (2023) comparing different machine learning algorithms, such as decision trees, random forests, and neural networks, considering their efficiency for processing diversified data types in stock market prediction. Similarly, Huang et al. (2019) contributed a CoStock model that integrates a Deep Factorization Machine (DeepFM) with attentional embeddings to capture complex feature interaction and temporal patterns in stock data.

3 Methodology

This study aims to the predict stock prices of Tesla (TSLA) and Amazon(AMZN), by combining sentiment analysis from social media with common indicators provided by technical Analysis. Three machine learning models - Long Short-Term Memory (LSTM) network, Random Forest, and Generative Adversarial Network (GAN), were created to achieve this purpose. These models were selected for their known effectiveness in handling time-series data and capturing intricate patterns with financial datasets (Mehta et al. (2021); Huang et al. (2023)). The methodology section of this research will describe the procedures, equipment, techniques used, and evaluation criteria.

3.1 Data Collection and Preprocessing

3.1.1 Data Sources

This research used the following data sources:

Stock Market Data: Historical stock prices for Tesla and Amazon were sourced from Yahoo Finance. This data contains important fundamental columns like, High, Low, Open, and Close Prices of individual organization stock with their respective Volumes. We selected Yahoo Finance as it provides the complete and most accurate information on historical financial data.

Social Media Data: Using tweets that mention Tesla and Amazon. This data was scraped from Twitter in real time to understand how people were talking about these companies. Twitter was chosen because it has a broad user base and huge availability of real-time data that reflects market sentiment Bollen et al. (2011)

3.1.2 Data Preprocessing

Stock Market Data: The stock data information was read from CSV files before carefully checking for any parsing errors in the output. In the case of error, the problematic lines were identified and corrected to ensure the integrity of the data. Only through this will the integrity of the data not be damaged; otherwise it cannot be reliably used for model training or evaluation.

The cleaned data in this project filtered out irrelevant information, keeping only certain columns of interest: Open, High, Low, Close, and Volume. Filtering is important to focus on the high-impact features of stock price prediction The following technical indicators have been computed and added to the data set: Moving Average (MA7, MA20), MACD, Bollinger Bands, Exponential Moving Average (EMA), Relative Strength Index (RSI), Williams %R, and Momentum. These are important indicators of various market dynamics and enhance the predictive powers of the models undertaken in this study Zhao et al. (2017) .

Social Media Data: Sentiment analysis on the gathered tweets was conducted using the NLTK VADER sentiment analyzer, whereby each tweet will be normalized and analyzed in computing sentiment scores (positive, negative, neutral, and compound).

This process helps to understand public sentiment, which is a major determinant factor in stock price changes Kumar and Ravi (2016). The sentiment scores were averaged daily and then merged with stock data on a date basis. The result of the integration will be a dataset that contains market and sentiment data, providing a holistic view of the factors influencing stock prices Chen et al. (2015).

3.1.3 Normalization

The combined dataset is then normalized using MinMaxScaler to scale features between -1 and 1. This normalization ensures consistency over all of its features and facilitates model training efficiency by improving the convergence rates and reducing training times. Normalization is a standard preprocessing step that helps handle diverse scales of data features Zhao et al. (2017).

3.2 Model Development

3.2.1 LSTM Model

LSTM networks are a type of Recurrent Neural Network (RNN) that can learn long-term dependencies in sequential data. The key components of an LSTM cell include the forget gate (f_t), input gate (i_t), cell state (C_t), and output gate (o_t) Hochreiter and Schmidhuber (1997). The equations governing these components are:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

where σ is the sigmoid function and tanh is the hyperbolic tangent function.

Data Preparation:

The sliding window approach was used to generate sequential data with a sequence length of 60 days. This helps to capture the temporal dependencies that exist in the data, which is essential for time series forecasting. The sliding window approach helps to create a dataset that LSTM can learn from effectively Fischer and Krauss (2018).

The dataset was split 80-20 into both training and test sets. This will ensure the model is trained a good portion of the data and is tested or evaluated on unseen data.

Model Architecture:

A Sequential model with two LSTM layers of 50 units each and a Dense output layer has been built. LSTM network is very powerful in capturing long-term dependencies from time-series data, which makes it suitable for stock price predictions Nelson et al. (2017).

The model was compiled using the 'adam' optimizer and the loss function 'mean_squared_error'. These are standard choices in many time-series forecasting applications to ensure efficiency during training and accuracy in predictions.

Model Training:

The model was then trained for 10 epochs, with a batch size of 32, and 10% of the training data used for validation. This validation helps in monitoring the performance of the model and preventing overfitting.

Model Evaluation:

Predictions were made on the test set, and results were compared against actual prices. The Root Mean Squared Error (RMSE) was used to evaluate the model's performance hence providing a measure of the model's accuracy.

3.2.2 Random Forest Model

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees to improve the predictive accuracy and control overfitting Breiman (2001).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{7}$$

where $T_i(x)$ is the prediction of the i-th tree and N is the total number of trees.

Data Preparation:

Features such as open, high, low, and volume, along with the target of Close from the dataset were scaled and then split into a train-test set as a Random Forest model benefits from feature scaling because it enhances its predictive performance Tiwari et al. (2023).

Model Architecture:

A random forest model with 100 estimators was constructed. Random Forests are a variety of ensemble learning and many decision trees are built during training, and the mean prediction of all the individual trees outputs the final prediction. The approach helps to avoid, to a large extent, overfitting problems and enhances the robustness, hence accuracy, in predictions; therefore, it is suitable for stock price forecasting Tiwari et al. (2023).

Model Training:

A RandomForestRegressor of 100 estimators was trained on the training set. Random Forests are known to be robust to overfitting and are certain to handle high-dimensional data efficiently.

Model Evaluation:

Predictions were then made against the test set, and these were compared to the actual prices. The model performance will be evaluated using the RMSE to easily measure prediction error. The accuracy, Precision, and Recall were also calculated for this model.

3.2.3 GAN Model

GANs consist of two neural networks, the generator G and the discriminator D, which compete against each other. The generator aims to produce realistic data, while the discriminator tries to distinguish between real and generated data Goodfellow et al. (2014). The loss functions for the generator and discriminator are:

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{8}$$

$$\mathcal{L}_G = -\mathbb{E}_{z \sim p_z(z)}[\log D(G(z))] \tag{9}$$

where x represents real data, z represents the noise vector, $p_{data}(x)$ is the data distribution, and $p_z(z)$ is the noise distribution.

Data Preparation:

The normalized dataset was used for the model also and the creation of sequential data was done using the sliding window approach. This approach will help in capturing the temporal dependencies in the data very effectively.

Model Architecture:

The generator model was built with a multiple of 5 LSTM layers with 3 dense layers as attention mechanisms that enable the capture of temporal dependencies. Attention mechanisms help in focusing on the most relevant parts of the input sequence and hence improve the accuracy of the prediction Chen et al. (2022).

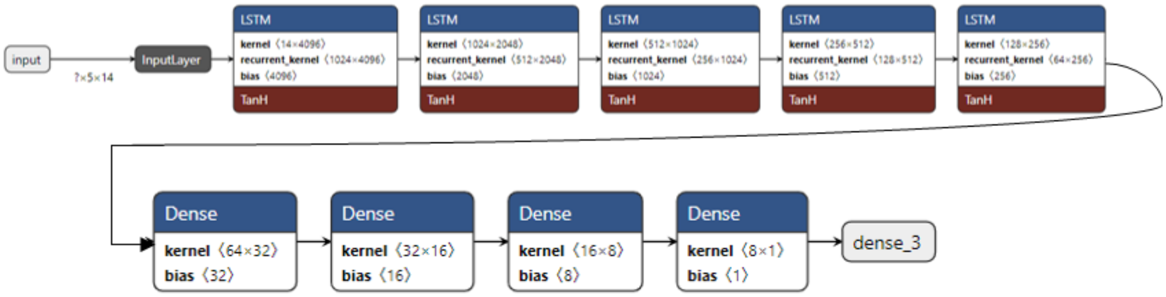

Figure 1: Detailed view of the Generator component

In the discriminator model, a multiple 5 convolutional was used, followed by 3 Dense layers for classifying real and fake data with a sigmoid activation function. This architecture makes the model very efficient at differentiating between real and synthetic data, ensuring the generation of high-quality predictions Zhao et al. (2017).

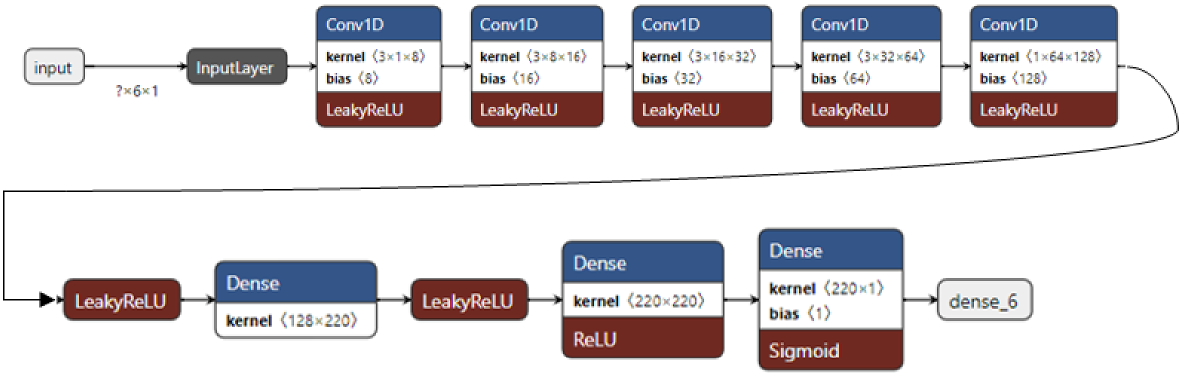

Figure 2: Detailed view of the Discriminator component

Training:

Models were trained with a self-defined training loop for 250 epochs, using a learning rate of 5e-4. This means that at each epoch, the generator generates synthetic stock prices, and the discriminator tells which one is real and which one is generated. This kind of adversarial training increases in quality over time Goodfellow et al. (2014).

Evaluation:

Predictions on the test set were obtained and matched with real prices. The accuracy, precision, and recall were calculated to evaluate the classification performance of the model. The RMSE was also calculated for regression performance, providing a comprehensive evaluation of the model Li et al. (2016).

3.3 Statistical Techniques and Analysis

3.3.1 Sentiment Analysis:

In this study, the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool was utilized as it attuned specifically to sentiments as expressed on social media platforms. This lexicon works quite well on short texts, typical in the case of Twitter posts, returning scores for positive, negative, and neutral sentiments with a compound score indicating the overall sentiment polarity Hutto and Gilbert (2014) and by integrating these sentiment scores with the traditional indicators of stock prices can enhance the model predictive power.

3.3.2 Feature Engineering

Feature engineering is a process of using domain knowledge to build features with the ability to make machine learning algorithms work better. In stock price prediction, feature engineering involves the development of meaningful inputs from raw financial data and sentiment scores Heaton et al. (2017). In this study, technical indicators were calculated to capture various parts of stock market movement to enhance the model's predicting abilities Zhao et al. (2017).

3.3.3 Model Performance Metrics:

Accuracy: The proportion of correct predictions among the total predictions.
Precision: The proportion of true positive predictions among the positive predictions.
Recall: The proportion of true positive predictions among the actual positives.
RMSE: A measure of the differences between predicted and actual values, providing a straightforward measure of prediction accuracy.

3.4 Equipment and Software

3.4.1 Hardware:

The models were trained on a high-performance computing system with sufficient computational resources (GPU/TPU) from Google Colab to handle deep learning tasks. This ensures efficient training and evaluation of the models.

3.4.2 Software

The following libraries and tools were used:

- Python for programming.
- Pandas and NumPy for data manipulation.
- Matplotlib for data visualization.
- TensorFlow and Keras for building and training neural networks.
- Scikit-learn for preprocessing and evaluation metrics.
- NLTK for sentiment analysis.

4 Design Specification

The primary goal of this study is to combine traditional financial data with social media sentiment analysis to predict future stock prices. To this purpose, three different models were considered: Random Forest, Long Short-Term Memory, and a Generative Adversarial Network. All the models capture different dimensions of information within the data with the GAN being the most advanced, including both financial indicators and sentiment analysis from social media.

4.1 Architectures of Models

4.1.1 Random Forest

Architecture: Random Forest is an ensemble learning method for classification and regression. During training, it constructs many decision trees and then outputs the average prediction or, in the case of classification, the mode of classes from individual trees. In this way, it manages the variance and bias trade-off very effectively Breiman (2001).

Functionality:

Feature Selection: In this study, the Random Forest model at any given time makes use of a subset of features in making every tree split. This helps in the strength of the model when it encounters high-dimensional data and issues of multicollinearity.

Ensemble Learning: By combining predictions from multiple trees, the Random Forest reduces the risk of overfitting and improves the model's generalizability.

Bootstrap Aggregation (Bagging): In this technique, every tree is trained on a random subset of data. This helps in reducing variance and making the model more stable, hence fitting for the diverse financial and sentiment features used in this study.

4.1.2 Long Short-Term Memory (LSTM)

Architecture: LSTM is a type of Recurrent Neural Network, which is extremely well-suited for time series prediction tasks because it is equipped with memory cells that hold information over time and gates to control the flow of information through the network Hochreiter and Schmidhuber (1997).

Functionality:

Memory Cells: In this study, memory cells are used by the model in LSTM networks to store information for long periods and to embed long dependencies within the stock price data.

Gates: The input, forget and output gates in LSTM cells control information flow, allowing the network to keep relevant information while erasing data that is irrelevant.

Sequential Data Handling: The LSTM model is well equipped to model the sequential nature of stock price data and effectively predicts future prices based on historical trends.

4.1.3 Generative Adversarial Network (GAN)

Architect: The GANs make use of the two neural mechanisms, one generating artificial data and the other testing for the authenticity of data; through adversarial processes, they are trained as partners. These mechanisms are referred to as the generator and the discriminator, respectively, and the discriminator makes the generator keep on improving the generated data through iteration Goodfellow et al. (2014).

Functionality:

Generator: In this study, the generator comprises of with five LSTM blocks followed by dense layers. The LSTM blocks capture the temporal dependency in time-series data, while the dense layers will help to refine the output into a near value of the real stock prices.

Discriminator: The discriminator consists of five convolutional layers and three dense layers, of which the last is activated with a sigmoid activation function. The convolutional layers extract features from input data, while the dense layers classify it as real or synthetic.

Adversarial Training: Since GANs are adversarial, they drive the generator to produce more realistic data over time, thus enhancing the power of the model to simulate complicated data distributions successfully.

4.2 Data Integration and Feature Engineering

4.2.1 Sentiment Analysis

Sentiment analysis involves the extraction and quantification of subjective information in social media to obtain a view regarding stock sentiment. Techniques such as VADER (Valence Aware Dictionary and Sentiment Reasoner) were used in analyzing sentiment from text data Hutto and Gilbert (2014). By integrating sentiment scores with financial data, the model is able to account for market sentiment and this can significantly influence stock prices.

For this study, VADER sentiment scores are extracted from the tweets concerning the stocks under review and to achieve this, the scores are then integrated with financial data to yield a complete dataset that includes market sentiment and traditional financial indicators.

4.2.2 Feature Engineering

The process of feature engineering involves taking raw data and transforming it into informative features that drive model performance. In this study, financial indicators such as moving averages, MACD, Bollinger Bands, RSI, Williams %R, and momentum have been computed. All these indicators capture the various aspects of stock price movements, aiding in making a comprehensive dataset for training the models.

Technical Indicators:

Moving Averages: Used to smooth out price data, providing a clearer picture of the trend direction by averaging the stock price over a specified period.

MACD (Moving Average Convergence Divergence): Highlights changes in the strength, direction, momentum, and duration of a stock's price trend.

Bollinger Bands: Provide a relative definition of high and low prices of a market, calculated based on the moving average and standard deviation.

RSI (Relative Strength Index): Measures the speed and change of price movements, used to identify overbought or oversold conditions.

Williams %R: A momentum indicator that measures overbought and oversold levels.

Momentum: Reflects the rate of change in stock prices, indicating the strength of price movements.

5 Implementation

In the implementation of this proposed solution, a step-by-step procedure was followed in a sequential manner ensuring the accuracy and efficiency of all models in the prediction of stock prices. The steps involved here are data collection, preparation and transformation, model development, and model Optimization.

5.1 Data Collection

The historical data for S&P 500 more specifically Tesla stock (TSLA) and Amazon stock (AMZN) were retrieved using the Yahoo Finance API.

It contained essential financial indicators like Open, High, Low, Close, and Volume for every day it trades. All the data was kept in a CSV file for further processing. In addition to this, the extraction of social media data from the Twitter API mentioning Tesla and Amazon has been retrieved. This will become a very critical step since the raw input is given to build and train predictive modeling.

5.2 Data Preparation and Transformation

The next step after the data collection was the data preparation and transformation and this began with Feature engineering which was done with Pandas and Numpy, providing many functions for the computation of various indicators. The key indicators that were computed for this implementation include Moving Averages MA7, MA20, MACD, Bollinger Bands, RSI, Williams %R, and Momentum, which capture the general trend in the market and its turning points.

Sentiment analysis was then performed with the NLTK VADER library by computing sentiment scores from each tweet obtained earlier. The tweets are normalized for consistency, their sentiment scores computed, and then averaged daily was obtained for merging with stock data. Any missing values in the data were replaced by zeros to maintain the integrity of the data. The data was normalized further to ensure consistency and comparability using the MinMaxScaler to keep the scaling features between –1 and 1. This normalization will deal with different scales of data and improve the performance of machine learning models.

5.3 Model Development

This study saw the developed three models: LSTM, Random Forest, and GAN. Each of these models was designed to leverage the strengths of the other models and provide a comprehensive evaluation of their predictive capabilities.

5.3.1 LSTM Model

The first model built in the course of this study was the LSTM model, developed with the Sequential API from TensorFlow's Keras library. The architecture here is built by an LSTM layer containing 50 units, succeeded by another similar layer with the same count of units, and finished with a Dense output layer. LSTM networks are particularly effective in capturing long-term dependencies in time-series data, making them suitable for stock price prediction Nelson et al. (2017). The model was compiled with an 'adam' optimizer and 'mean_squared_error' loss function, standard choices for time-series forecasting. Training will consist of the actual fitting of the model on previously normalized data of stock prices to predict future prices.

5.3.2 Random Forest Model

The Random Forest model utilizes the RandomForestRegressor of the Scikit-learn library. In developing this model, 100 trees or estimators were used, a common choice for balancing model complexity and performance Breiman (2001). Random Forests are known to be very robust and have the ability to handle high-dimensional data without any fear of overfitting. It was trained on the same feature set used in the LSTM model, including historic stock prices, technical indicators, and sentiment scores. During training, the dataset was divided into a training set and a test set to evaluate the model's performance and further tune the hyperparameters for the best possible accuracy.

5.3.3 GAN Model

The architecture of the GAN model consisted of one generator and a discriminator. The generator was designed using five LSTM blocks (units: 256, 128, 64, 32, 16) for capturing the temporal dependencies of the stock prices. The increased number of LSTM units and layers in the generator can allow for learning of complex patterns and dependencies in time-series data regarding stock prices. After each LSTM layer, dropout layers were put together to avoid over-fitting and help in generalization. Finally, the generator output amounted to a synthetic series of stock prices, which would, in turn, be evaluated by the discriminator.

The discriminator included five convolutional layers and three dense layers with sigmoid activation functions to enhance the ability of the discriminator to differentiate between real and fake sequences. The convolutional layers were to extract features from the input sequences, and the dense layers, equipped with enhanced activation functions, provided better classification capability. All of these structures helped the discriminator to be able to differentiate between real stock price sequences from the generated ones and thus formed an effective adversarial training process.

5.4 Model Optimization

Optimization and tuning of the models were done iteratively with different modifications and evaluations to increase performance. Key changes and improvements have been made regarding model architecture and training procedures.

5.4.1 Generator Model Enhancements

The generator model in GAN was improved by significantly increasing the number of units in each LSTM layer (e.g., 1024, 512, 256, 128, 64). This enhancement allowed the generator to capture temporal dependencies in the stock price data more effectively. Another adjustment is the addition of more LSTM layers to the generator model. This deeper architecture helps the model learn more complex patterns in the data.

5.4.2 Discriminator Model Improvements

The Discriminator Model was improved further increasing the number of Conv1D layers in the discriminator model, allowing for better feature extraction from the input sequences. Also, more dense layers were introduced with LeakyReLU activation in this case for enhancing the capturing and ascertaining of real or fake sequences.

5.4.3 Training Procedure Adjustments

Training procedures were fine-tuned to enhance the optimization process. Improved gradient tape handling using TensorFlow's gradient tape ensured better optimization of the loss functions for both generator and discriminator models. The checkpointing intervals were increased, which provided better performance monitoring and rollback if necessary.

5.4.4 Summary of Key Changes

- Increased LSTM units and layers in the generator improved the ability to capture temporal dependencies.
- Additional Conv1D layers in the discriminator enhanced feature extraction from input sequences.
- Dense Layers Enhancement in the discriminator improved the ability to distinguish between real and fake sequences.
- Improved Gradient Tape Handling ensured better optimization of loss functions.
- Frequent Checkpointing allowed better monitoring of model performance and rollback if necessary.

6 Evaluation

This section analyzes the main findings and results regarding predicting stock prices using different machine learning models: Random Forest, LSTM, and GAN. By examining the performance of these models in predicting stock prices for TSLA and AMZN, impactful insights can be provided from both academic and practitioner perspectives, highlighting the most relevant results that align with this study's research questions and objectives, and critically evaluating the experimental outputs. Statistical tools were employed to analyze the significance of the findings critically and, in this way, give deep insights into the effectiveness and robustness of developed models. Evaluation will also pinpoint the strengths and weaknesses of each model and will also indicate opportunities for further research and practical application in forecasting finance.

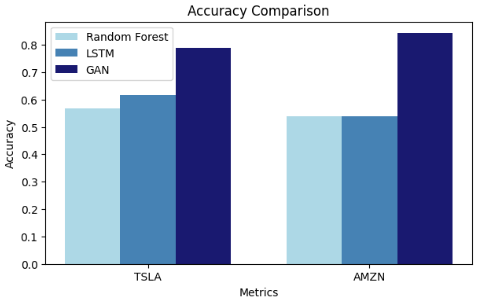

Figure 3: Comparison of Metrics for Accuracy

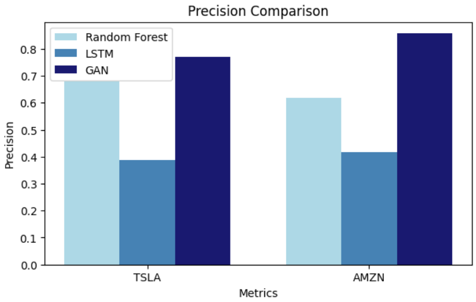

Figure 4: Comparison of Metrics for Precision

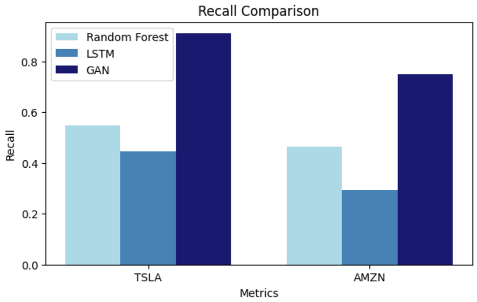

Figure 5: Comparison of Metrics for Recall

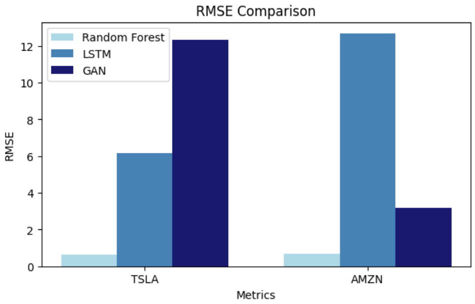

Figure 6: Comparison of Metrics for RMSE

6.1 Experiment 1: Random Forest Model Evaluation

The random forest model performed moderately for TSLA and AMZN. The precision was quite high for the case of TSLA, in comparison with that of AMZN, meaning a lot of real positives had been identified. However, the overall accuracy or recall depicts that there have to be some limitations of this model in predicting the correct movement of stock.

Table 1: Random Forest Model Performance Results for TSLA and AMZN Stocks

Performance Metric	Stock	Random Forest Results
RMSE	TSLA	0.6567
	AMZN	0.6782
Accuracy	TSLA	56.86%
	AMZN	54.00%
Precision	TSLA	68.0%
	AMZN	61.91%
Recall	TSLA	54.8%
	AMZN	46.42%

6.2 Experiment 2: LSTM Model Evaluation

The LSTM model performed more efficiently in terms of capturing the long-term dependencies in time-series data and hence showed better accuracy for TSLA compared to the Random Forest model. The values, however, for precision and recall were pretty low, thereby showing that it was quite hard to correctly identify the positive cases. For AMZN, the model did very poorly with high RMSE, which means significant prediction errors.

Table 2: LSTM Model Performance Results for TSLA and AMZN Stocks

Performance Metric	Stock	LSTM Model Results
RMSE	TSLA	6.153
	AMZN	12.667
Accuracy	TSLA	61.53%
	AMZN	53.84%
Precision	TSLA	38.9%
	AMZN	41.66%
Recall	TSLA	44.44%
	AMZN	29.41%

6.3 Experiment 3: GAN Model Evaluation

The GAN model performed much better compared with both the Random Forest and LSTM models. The GAN showed the highest accuracy, precision, and recall on TSLA; thus, it worked well for predicting the movement of this stock. It also worked fine on AMZN for the accuracy and precision values. The reduced RMSE value in the case of AMZN meant more reliable predictions as realized on TSLA.

Table 3: GAN Model Performance Results for TSLA and AMZN Stocks

Performance Metric	Stock	GAN Model Results
RMSE	TSLA	12.35
	AMZN	3.180
Accuracy	TSLA	78.94%
	AMZN	84.21%
Precision	TSLA	76.92%
	AMZN	85.714%
Recall	TSLA	90.90%
	AMZN	75.00%

6.4 Discussion

The evaluation of the models reveals the fact that the GAN model consistently outperforms the Random Forest and the LSTM models for both TSLA and AMZN stocks in terms of accuracy, precision, and recall. This aligns with the literature, which highlights the superior ability of GANs to capture complex patterns in time-series data.

Although the Random Forest model is useful for its interpretability and simplicity, it, however, would perform poorly due to stock price movements being very volatile and non-linear. While the LSTM model is arguably the best model for such a problem, it too faced challenges in predicting stock prices accurately, particularly for AMZN. This can be due to the stock market has inherent noise and the unpredictable nature of its data, which may alter its learning behavior.

The adversarial training mechanism of the GAN model makes it successful in generating more realistic and accurate predictions. The generator's use of LSTM blocks helps capture temporal dependencies through LSTM blocks, while the discriminator impeccably differentiates between real and generated data, improving the model's robustness and accuracy.

7 Conclusion and Future Work

7.1 Conclusion

The pivotal question that drove this research was the ascertain the extent to which deep learning models, mostly GANs, can predict the stock market through analysis of social media trends and sentiments captured from X (formerly Twitter). This study focused on comparing the predictive efficiency of GANs against traditional models like RF and LSTM networks using datasets obtained from TSLA and AMZN stocks.

The result indicated that, on most performance metrics, GAN models performed significantly better than both RF and LSTM models. For example, the GAN model had an accuracy of 78.94% for TSLA and 84.21% for AMZN. Corresponding precision rates were 76.92% and 85.71%, while recall had rates of 90.90% and 75.00%, respectively. These results underscore the advanced capability of GANs in distinguishing and learning intricate dynamics and temporal correlations within stock price data driven by insights from social media.

The GAN model had a very strong performance, which underscores the view that it could quite easily become a powerful tool in financial forecasting, with improved precision and reliability over traditional approaches. This supports the notion that the integration of deep learning with social media analytics can significantly enhance stock market prediction.

However, the study is not without its limitations as the high computational demand for training GANs and strong dependency on extensive good-quality input data set a big challenge. Further, the model used in the study was applied only to two classes of stocks: TSLA and AMZN; necessitating further studies to explore its efficacy across diverse market scenarios and stock categories.

7.2 Future Work

Looking ahead, expanding the scope to encompass more stocks and market indices might shed more light on how adaptive and effective the GAN model is across different sectors of the economy. Increasing the enrichment of model inputs by macroeconomic factors and sector-specific news may further enhance predictive accuracy.

Improvements to sentiment analysis, to include more complex natural language processing techniques such as BERT or GPT-3, could give more nuanced interpretations of market sentiments and enhance the predictive quality of the models.

Another area of experimentation could be other GAN architectures, such as Wasserstein GANs or Conditional GANs, that position themselves to continue to optimize predictive performance. Practical application testing in real-time market conditions would be invaluable to verify the model's effectiveness in live trading environments.

Addressing the computational intensity of GAN training could be of immense significance and using new advanced solutions like distributed computing or parallel processing allows these models to be scaled up efficiently. This would do a great job of offering more options for really extended and fast model train phases, lending a way toward real-time analytics and commercial applications in this field of financial forecasting.

References

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market, *Journal of computational science* **2**(1): 1–8.

Breiman, L. (2001). Random forests, *Machine learning* **45**: 5–32.

Chandola, D., Mehta, A., Singh, S., Tikkiwal, V. A. and Agrawal, H. (2023). Forecasting directional movement of stock prices using deep learning, *Annals of Data Science* **10**(5): 1361–1378.

Chen, K., Zhou, Y. and Dai, F. (2015). A lstm-based method for stock returns prediction: A case study of china stock market, *2015 IEEE international conference on big data (big data)*, IEEE, pp. 2823–2824.

Chen, R.-C., Yang, W.-I. and Chiu, K.-C. (2022). Transfer learning and lstm to predict stock price, *2022 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE, pp. 165–169.

Chopra, R. and Sharma, G. D. (2021). Application of artificial intelligence in stock market forecasting: a critique, review, and research agenda, *Journal of risk and financial management* **14**(11): 526.

Fischer, T. and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions, *European journal of operational research* **270**(2): 654–669.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets, *Advances in neural information processing systems* **27**.

Heaton, J. B., Polson, N. G. and Witte, J. H. (2017). Deep learning for finance: deep portfolios, *Applied Stochastic Models in Business and Industry* **33**(1): 3–12.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory neural computation 9 (8): 1735–1780, *Search in* .

Huang, J.-Y., Tung, C.-L. and Lin, W.-Z. (2023). Using social network sentiment analysis and genetic algorithm to improve the stock prediction accuracy of the deep learning-based approach, *International Journal of Computational Intelligence Systems* **16**(1): 93.

Huang, J., Zhang, X. and Fang, B. (2019). Costock: a deepfm model for stock market prediction with attentional embeddings, *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 5522–5531.

Huang, Z., Lin, Y. and Xue, H. (2022). A hybrid model combined deep learning approaches in stock price prediction, *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, IEEE, pp. 835–838.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the international AAAI conference on web and social media*, Vol. 8, pp. 216–225.

Kumar, B. S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain, *Knowledge-Based Systems* **114**: 128–147.

Kumar, R., Sharma, C. M., Chariar, V. M., Hooda, S. and Beri, R. (2022). Emotion analysis of news and social media text for stock price prediction using svm-lstm-gru composite model, *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, IEEE, pp. 329–333.

Lakshmanarao, A., Babu, M. R., Gupta, C. and Lakshmi, A. S. G. (2022). Stock price prediction using deep learning and flask, *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, pp. 1–5.

Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., Min, H. and Deng, X. (2016). Empirical analysis: stock market prediction via extreme learning machine, *Neural Computing and Applications* **27**: 67–78.

Lim, M. and Yeo, C. K. (2020). Harvesting social media sentiments for stock index prediction, *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*, IEEE, pp. 1–4.

Mehta, P., Pandya, S. and Kotecha, K. (2021). Harvesting social media sentiment analysis to enhance stock market prediction using deep learning, *PeerJ Computer Science* **7**: e476.

Nelson, D. M., Pereira, A. C. and De Oliveira, R. A. (2017). Stock market's price movement prediction with lstm neural networks, *2017 International joint conference on neural networks (IJCNN)*, Ieee, pp. 1419–1426.

Nijaguna, G. (2023). A time weighted based deep learning model for stock market price prediction, *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, IEEE, pp. 1–5.

Pang, X., Zhou, Y., Wang, P., Lin, W. and Chang, V. (2020). An innovative neural network approach for stock market prediction, *The Journal of Supercomputing* **76**: 2098–2118.

Rao, T. and Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications, *Knowledge-based systems* **89**: 14–46.

Sonkiya, P., Bajpai, V. and Bansal, A. (2021). Stock price prediction using bert and gan, *arXiv preprint arXiv:2107.09055* .

Sutradhar, K., Sutradhar, S., Jhimel, I. A., Gupta, S. K. and Khan, M. M. (2021). Stock market prediction using recurrent neural network's lstm architecture, *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 0541–0547.

Tiwari, V., Lohani, B. P., Rana, A., Pandey, U. P. and Dhariwal, B. (2023). Stock market prediction using different machine learning algorithms, *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Vol. 10, IEEE, pp. 147–151.

Wang, Y. (2023). Deep learning-based techniques for influencing and predicting the impact of public figures' social media statements on the stock market, *2023 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, IEEE, pp. 93–96.

Zhang, X., Zhang, L., Xu, L. and Jiang, Y. (2022). Research on influential factors in stock market prediction with lstm, *2022 7th International Conference on Big Data Analytics (ICBDA)*, IEEE, pp. 25–29.

Zhao, Z., Rao, R., Tu, S. and Shi, J. (2017). Time-weighted lstm model with redefined labeling for stock trend prediction, *2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI)*, IEEE, pp. 1210–1217.