Strategic Portfolio Management: Analyzing Netflix and Apple Stocks

Sneed Murillo, Shayak Chaudhuri, An Dang

Georgia Institute of Technology

Abstract

This study explores the predictive power of sentiment analysis on stock price movements, focusing on Netflix and Apple stocks from July 2023 to July 2024. By integrating sentiment data from both traditional media and social media platforms, particularly Reddit, we aim to build a comprehensive predictive model. The study is structured in two stages: first, we perform sentiment analysis using VADER and FinBERT models to classify sentiments from Reddit posts and news articles. Second, we employ machine learning algorithms including LSTM, GRU, GAN, XG Boost, and Random Forest Regressor to predict stock prices based on the sentiment scores. Our findings indicate that while traditional sentiment analysis tools offer valuable insights, they require significant fine-tuning to capture the nuances of investor sentiment, particularly on social media platforms. The study highlights the potential of combining diverse sentiment sources to enhance stock price prediction, providing actionable insights for retail investors.

Index Terms – Financial markets, sentiment analysis, investor sentiment, stock market behavior, predictive modeling.

TABLE OF CONTENTS

PROJECT OVERVIEW ***3***

INTRODUCTION ***3***

LITERATURE REVIEW ***4***

Efficient Market Hypothesis (EMH) 4

Behavioral Finance 4

Sentiment Analysis in Finance 4

Machine Learning Models in Financial Forecasting 5

DATA OVERVIEW ***5***

Data Collection **5**

Stock Data 5

Social Media Posts 6

Financial News 6

Data Preprocessing **6**

MODELING OVERVIEW ***7***

Sentiment Analysis **7**

Time Series Analysis **9**

LSTM (Long Short-Term Memory) 9

GRU (Gated Recurrent Units) 9

GAN (Generative Adversarial Network) 9

XG Boost (Extreme Gradient Boosting) 10

Random Forest Regressor 10

Evaluation Metric 10

RESULTS ***11***

Sentiment Analysis **11**

Time Series Analysis **14**

CHALLENGES ***20***

Data Collection Limitations 20

Data Quality 20

Methodological Adjustments 20

Model Performance 20

Manual Labeling and Validation 21

Data Integration 21

CONCLUSION ***21***

REFERENCES ***22***

PROJECT OVERVIEW

This project represents a synergy of specialized skills and knowledge. Each team member contributed with a unique and valuable perspective according to their expertise, allowing us to solve problems efficiently and adopt a multidimensional approach. The collaboration among team members ensured that the project was executed comprehensively, combining insights from finance, project management, and data analysis.

The team members include:

- Sneed Murillo: With expertise in project management and finance, Sneed led the project by coordinating tasks, managing timelines, and ensuring all objectives were met. Sneed's financial analysis skills were vital in interpreting stock data and providing strategic insights. Additionally, Sneed authored the code for the project's initial sentiment analysis phase.
- Shayak Chaudhuri: With a background in Visual Analytics and Research, Shayak managed code validation and the overall interpretation and evaluation of the project's results. Shayak's expertise in debugging and optimizing code ensured accurate model implementation and performance. Shayak validated the code written by Sneed and An and played a significant role in interpreting and evaluating the outcomes.
- An Dang: An, specializing in business analytics, implemented the time series analysis models and ensured prediction accuracy. An's problem solving skills were crucial for developing robust models to analyze stock price movements based on sentiment scores. An also developed and trained the predictive models.

INTRODUCTION

"Investor sentiment, captured through various media, can offer deep insights into future market movements" (Tetlock, 2007). This study aims to predict stock prices using sentiment scores derived from both traditional media and social media posts from Reddit. Building on previous research that primarily focuses on a single media source (Fan & Gordon, 2014; Ma et al., 2020; Mittal et al., 2021), this work integrates data from multiple influential factors to develop a comprehensive predictive model for stock price movements.

Traditional media have long played a crucial role in shaping investor perceptions and driving market movements. For instance, García-Méndez et al. (2022) demonstrated how media bias influenced stock markets during the COVID-19 pandemic, highlighting the significant impact of traditional media on portfolio management. Similarly, Li and Pan (2022) underscored the strong correlation between financial news and stock price fluctuations.

In recent years, social media has emerged as a powerful force in financial markets. Platforms such as Reddit's r/wallstreetbets have become central to retail investing, where individual investors share ideas, discuss strategies, and collectively influence stock prices. Hasso et al. (2022) examined speculative activities on r/wallstreetbets, particularly its impact on Bitcoin, and highlighted the forum's capacity to drive market movements. The dramatic rise of GameStop in 2021, fueled by this subreddit, further exemplifies the substantial influence of online discussions on stock prices (Hu et al., 2021).

Initially, the focus of this study was classifying stock purchase decisions. However, due to the lack of specific stock-related data on r/wallstreetbets, we expanded our scope to include other popular subreddits such as r/stocks, r/investing, r/Daytrading, and r/StockMarket. This change allowed us to gather a more robust dataset, including posts about major stocks like Apple and Netflix. Additionally, we incorporated information derived from emojis, recognizing their significant contribution to sentiment analysis.

The study was structured in two main stages. In the first stage, sentiment analysis was conducted using data collected from Reddit and news articles. In the second stage, a time series regression was performed to predict stock prices based on the obtained sentiment scores. This dual approach provides a more comprehensive understanding of how sentiments expressed across different platforms collectively influence stock market behavior. By building a model that captures the spread between market sentiment from online newspapers and public sentiment from Reddit, one can predict the stock market price on day T given the historical closing prices, news headlines, and Reddit posts in the period T0 – T1.

LITERATURE REVIEW

Efficient Market Hypothesis (EMH)

The Efficient Market Hypothesis (EMH) posits that stock prices fully reflect all available information at any given time. Introduced by Eugene Fama in the 1960s, EMH implies that it is impossible to consistently achieve higher returns than average market returns on a risk-adjusted basis, as stock prices should only respond to new, random, and unpredictable information (Fama, 1970). EMH is categorized into three forms: weak, semi-strong, and strong. Weak Form Efficiency asserts that all past trading information is already reflected in stock prices, suggesting that technical analysis is ineffective. Semi-Strong Form Efficiency claims that all publicly available information is reflected in stock prices, making both technical and fundamental analysis futile. Strong Form Efficiency argues that all information, including insider information, is fully reflected in stock prices.

The relevance of EMH to this study lies in understanding the limitations and potential predictability of stock prices based on available information from traditional news outlets and social media platforms. If markets are not perfectly efficient, leveraging sentiment from these sources could provide actionable insights for predicting stock price movements.

Behavioral Finance

Behavioral Finance challenges EMH by incorporating psychological theories into finance, explaining why investors might not always act rationally. Influenced by Daniel Kahneman and Amos Tversky, this field explores how cognitive biases, emotions, and social factors impact investors' decisions and market outcomes (Kahneman & Tversky, 1979). Overconfidence Bias leads investors to overestimate their knowledge and trading abilities, increasing market volatility. Herd Behavior causes individuals to mimic the actions of a larger group, often leading to irrational market trends and bubbles. Anchoring sees investors relying too heavily on initial information, even if irrelevant. Loss Aversion makes investors more sensitive to losses than gains, often resulting in risk-averse behavior.

By examining sentiment from r/wallstreetbets and traditional news outlets, this study aligns with behavioral finance principles, acknowledging that market movements can be influenced by irrational investor behavior driven by emotions and social interactions.

Sentiment Analysis in Finance

Sentiment analysis, or opinion mining, uses natural language processing (NLP) and machine learning techniques to identify and quantify sentiment expressed in text data. In finance, sentiment analysis gauges market sentiment and its potential impact on stock prices, theorizing that positive sentiment can drive stock prices up, while negative sentiment can lead to declines. VADER (Valence Aware Dictionary

and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically attuned to social media contexts, assigning sentiment scores based on a dictionary of sentiment-related words and phrases (Hutto & Gilbert, 2014). FinBERT, a pre-trained language model tailored for financial communication sentiment analysis, builds on the BERT model, fine-tuned with financial texts to accurately capture sentiment nuances in financial news articles (Yang et al., 2020).

By leveraging VADER and FinBERT, this study aims to quantify sentiment from Reddit posts and financial news, respectively, providing a comprehensive view of market sentiment from diverse sources. This analysis supports the hypothesis that combining sentiments from multiple media sources can enhance the prediction of stock price movements.

Machine Learning Models in Financial Forecasting

Machine learning models are increasingly used in financial forecasting due to their ability to identify patterns and relationships in large datasets that are not easily discernible through traditional statistical methods. This study applies several supervised learning algorithms, each with distinct theoretical foundations. The application of these models in this study aims to evaluate their performance in predicting stock price movements based on sentiment analysis, identifying the most effective approach for financial decision-making.

DATA OVERVIEW

Data Collection

For the development of the sentiment analysis of Netflix and Apple, three data sources were used: stock data, social media posts, and financial news. The latter two were used because of the influence of investors' decision-making. According to Fan and Gordon (2014), 80% of investors' decisions are influenced by news and social media discussions. These sources allow us to evaluate how general market sentiment influences stock prices. By combining quantitative analysis of stock data with qualitative analysis of sentiments derived from social media and news, this approach delivers a more comprehensive and accurate view of the factors influencing stock price behavior.

Stock Data

We used historical stock data for Netflix (NFLX) and Apple (AAPL) due to these companies having considerable influence on the market. They also have a constant presence on social media discussion forums and news headlines.

- Netflix: As a leader in the streaming industry, Netflix has revolutionized how we consume content. Its cultural and commercial impact is significant, making it an interesting subject for sentiment analysis. A recent study by Hu et al. (2021) shows how social media mentions can influence the stock prices of highly media-exposed companies like Netflix.
- Apple: As one of the most valuable companies in the world and a leader in technological innovation, fluctuations in Apple's stock price affect not only its investors but also the market. According to an article by Ma et al. (2020), discussions about Apple on social media platforms can reliably predict movements of the ticker prices due to the company’s large following and brand loyalty.

The stock prices were obtained through the Yfinance library in Python, which allows downloading historical financial data from Yahoo Finance. The downloaded data included adjusted closing prices, trading volumes, and other financial indicators from January 1, 2024, to May 31, 2024. Netflix is a less popular stock compared to Apple. The data available for Netflix is half of that of Apple. This imbalance creates another interesting benchmark to evaluate the performance of the sentiment analysis and time series prediction.

Social Media Posts

According to a study by Hasso et al. (2022), discussions on r/wallstreetbets can significantly impact stock prices and market volatility. Reddit, as a platform where opinions are shared more freely and with less moderation than traditional media, provides a more authentic view of investor sentiment.

To capture this perspective comprehensively, we expanded our data collection scope to include other prominent investing subreddits such as stocks, investing, Daytrading, and StockMarket. This broader approach enabled us to compile a more extensive and diverse dataset.

We utilized the PRAW library in Python to extract recent posts from these subreddits. The data extraction process involved authenticating and accessing the Reddit API to collect post titles, scores, IDs, URLs, and creation dates. This step was crucial as it provided unique insights into market movements influenced by real-time investor sentiment, thereby capturing a wide array of opinions and discussions that extend beyond traditional financial fundamentals.

Financial News

Financial news was collected using three different APIs: Currents API, Mediastack API, and News API, to filter relevant news on finance, stocks, and markets from recognized sources. This approach helps mitigate bias and provides a cohesive view of the market. Li and Pan (2022) recommend using a variety of news sources to avoid biases and obtain a comprehensive market view. García-Méndez et al. (2022) also emphasize the importance of collecting news from multiple platforms to avoid the bias of relying on a sole source. To further enhance the robustness of our dataset, we expanded our scope by incorporating GNews, which provided an additional layer of diverse and comprehensive financial news coverage. This expansion ensured that we captured a wider array of news articles, contributing to a more balanced and representative data set.

Using these APIs, we fetched news articles related to Apple and Netflix from various sources. The details of the data collection include keywords such as 'Apple', 'Netflix', 'AAPL', and 'NFLX', covering the last twelve months. The data collected includes the title, description, URL, and published date of each article.

Data Preprocessing

For our analysis, the data will be preprocessed to ensure accuracy and consistency. After collecting the data, various methods were applied to clean and process the datasets for sentiment analysis and to study the correlation between financial news, social media posts, and ticker prices.

First, there are several steps involved in preprocessing historical stock data for Netflix (NFLX) and Apple (AAPL). The downloaded data included adjusted closing prices, trading volumes, and other financial indicators. To prepare this data for analysis, we calculated daily stock returns by taking the percentage change in adjusted closing prices from one day to the next. Any missing values in the data were addressed using forward and backward filling methods, ensuring continuity and completeness. Additionally, outliers that could potentially distort the analysis were identified and removed to maintain the integrity and representativeness of the dataset.

The social media posts from the subreddits underwent a thorough cleaning and normalization process. As mentioned above, we extracted relevant posts and gathered post titles, scores, IDs, URLs, and creation dates using the Python Praw library. Each post's text was then processed to normalize the content,

which involved converting text to lowercase, removing punctuation and numbers, and eliminating stopwords using the NLTK library.

For the financial news articles collected, a similar cleaning and normalization process was applied. To ensure consistency across the dataset, the text from the news articles was converted to lowercase, punctuation and numbers were removed, and stopwords were eliminated.

By meticulously processing the data from these diverse sources, we ensured that the datasets were accurate, consistent, and suitable for in-depth analysis.

MODELING OVERVIEW

Sentiment Analysis

Sentiment analysis commonly classifies human messages under three categories: positive, negative, and neutral. In the context of financial analysis, these classifications can be adjusted to represent investment decisions:

- Bear (-1): the sentiment toward the stock is negative, it is time to sell
- Hold (0): the sentiment toward the stock is neutral, one should hold or do nothing
- Bull (1): the sentiment toward the stock is positive, it is time to invest

For Reddit sentiment analysis, we use the pretrained Vader model (short for V*alence Aware Dictionary & Sentiment Reasoner*), an open-source algorithm that is specifically built for social media analysis. The model should correctly classify the sentiment expressed in the following Reddit posts based on their titles:

- *"Netflix per capita revenue increases North America"* => positive or 1
- *"Nflx earnings thread"* => neutral or 0
- *"berkshire hathaway cuts aapl position"* => negative or –1

However, consider these posts with added Reddit flavors below:

- *"k aapl apple yolo tim cook pls cook tendies"* – using slang *"tendies"*
- *"aapl little yolo"* – using slang *"yolo"*
- *"well played mr tim apple"* – using sarcasm
- *"aapl good short"* – while *"good"* is positive, shorting is negative
- *"aapl amzn gains"* – using the abbreviated word "*amzn*" for "amazing"

While Vader excels in interpreting the informal and often colloquial language found on platforms, it misclassifies the examples above due to the complexities in understanding the r/wallstreetbets lingo. This task can even be challenging for humans who are not familiar with this online discussion forum. To improve the performance of the model, innovative words and associated scores will be added to the vocabulary. additional fine-tuning is required to account for the unique Reddit flair, such as memes, sarcasm, and slang, ensuring more accurate sentiment detection (Hutto & Gilbert, 2014)

For the news analysis, we use the Hugging Face model FinBERT. This model is pretrained to do natural language processing analysis on financial corpus. Although we anticipate the model to perform well, it is still important to evaluate its performance to see if additional fine-tuning is needed (Yang et al., 2020).

The results of sentiment analysis can be directly applied to financial decision-making processes. For example, investment firms can use sentiment scores to adjust their trading strategies, responding quickly to shifts in public opinion. Additionally, companies might monitor sentiment to gauge market reaction to their news releases or to anticipate stock price volatility. Despite its usefulness, sentiment analysis is not always reliable. Models like VADER and FinBERT require continuous training to maintain their performance, especially in the rapidly changing landscape of social media and news.

We will use F1 to compare the performance of the models. The model with the highest F1 score is considered the best in balancing precision and recall. In this case, we can define each evaluation metric as follows:

- True Positives (TP): Correctly predicted stock price increases.
- True Negatives (TN): Correctly predicted stock price decreases.
- False Positives (FP): Predicted stock price increases that fell, leading to monetary loss.
- False Negatives (FN): Predicted stock price decreases that rose, resulting in missed potential gains.

False positives indicate financial losses because the prediction suggested a price increase that did not materialize, leading to poor investment decisions. Conversely, false negatives represent missed profit opportunities, as the stock predicted to fall rise in value. Different investment strategies have distinct levels of risk tolerance so we can judge the effectiveness of the models accordingly.

The formulas for precision, accuracy, recall, and F1-score are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

To evaluate our models rigorously, we will employ k-fold cross-validation using k equals 5. This method involves splitting the dataset into five subsets, randomly choosing four subsets to train the model, and testing it on the last subset. We repeat this iteration five times so each subset is used to validate once. The final evaluation measurement will be the average of the metrics from each k fold, which provides a reliable assessment of the model performance.

Prior to these calculations, manual labeling was employed to generate the true sentiment labels. This manual labeling was applied to a representative sample of the data, providing a reliable basis for comparison. Based on these true sentiment labels, precision, recall, and F1-Score were calculated for each sentiment category, along with the confusion matrix. The confusion matrix shows the distribution of errors across different classes, providing insights into model performance.

Scenario Analysis:

- **High Accuracy but Low F1 Score**: If a model shows high accuracy but low F1 score, it suggests that the model is correctly predicting a large number of instances but might be doing so predominantly for the majority class. This indicates deficient performance in the minority class, with either low precision, low recall, or both.
- **High F1 Score but Lower Accuracy**: A model with a high F1 score but lower accuracy is likely to perform well in terms of precision and recall for the positive class, even if it makes more mistakes overall. This situation is preferable in contexts where correctly identifying the minority class is crucial, such as in financial risk analysis.

Given the nature of our data and the importance of correctly identifying minority classes (e.g., negative sentiments in financial news), we prioritize the F1 score over accuracy. The F1 score provides a better measure of a model's effectiveness in scenarios where false positives and false negatives have significant consequences. While we consider accuracy to ensure overall performance, our primary criterion for model selection is the F1 score due to its ability to capture both precision and recall effectively.

Time Series Analysis

We decided to frame the stock prediction problem as a multivariate time series analysis due to the sequential nature of the data and the numbers of features available. The following models were used:

LSTM (Long Short-Term Memory)

This model was chosen because it performs well with sequential time series data. Unlike the regular Recurrent Neural Network (RNN) model, LSTM fixes the problem of exploding and vanishing problems by adding a forget mechanism on top of the input and output gates. The neural network can also handle multiple input features and model their long-term dependencies. *(Olah, 2015)*

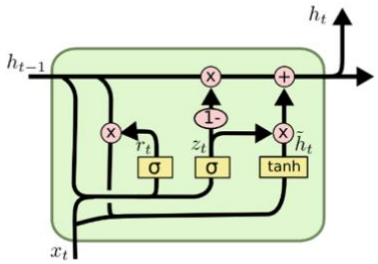


Figure 1. *LSTM Architecture (Olah, 2015)*

GRU (Gated Recurrent Units)

GRU is another gated RNN model that performs well on sequential data. Unlike LSTM, GRU's architecture only consists of two types of gates: update and reset. Due to having fewer hyperparameters to train, GRU is faster than LSTM. *(Sonkiya, 2021)*

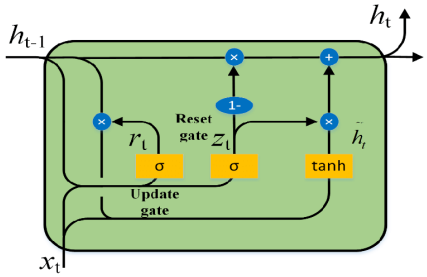


Figure 2. *GRU Architecture (Li, 2020)*

GAN (Generative Adversarial Network)

GAN was originally designed to generate realistic synthetic data, which makes it useful for improving predictions by creating additional training examples. This feature is especially useful in scenarios where labeled data is scarce. The model was chosen because it performs well with sequential time series data and

can effectively capture complex patterns in stock price movements. *(Sonkiya, 2021)* Moreover, since we have a small dataset, this is a suitable model to test.

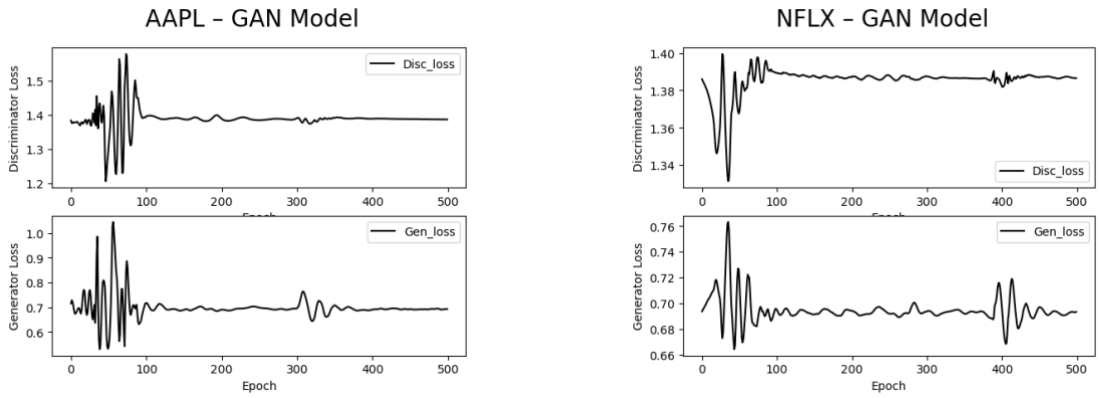


Figure 3. *GAN discriminator generator training loss over epochs*

In this scenario, the GAN model's training loss plots indicate that both the discriminator and generator experience significant fluctuations initially, reflecting the adversarial learning process. After around 100 epochs, both losses stabilize, with the discriminator loss around 1.3 and the generator loss around 0.7. This stabilization suggests that the model has achieved a reasonable balance, with the discriminator effectively distinguishing reality from generated data and the generator improving its ability to produce convincing data. The convergence of these losses implies that the training process for stock price prediction proceeds well.

XG Boost (Extreme Gradient Boosting)

XG Boost leverages stochastic gradient boosting mechanism for time series data. This machine learning model can effectively handle outliers, overfitting, and multiple features in large datasets. In recent years, there have been many attempts to utilize XG Boost for investment portfolio management. *(Chen, 2021)*

Random Forest Regressor

Random Forest is another tree-based model built on majority voting and random sampling. Parallel computing can be used to train multiple decision trees simultaneously to reduce training time, making the model effective in handling long time series. It can also manage multiple predictors and capture non-linear relationships. Despite being a simpler model compared to neural networks, Random Forest offers better interpretability. *(Sonkiya, 2021)*

Other models such as ARIMA were considered. However, they are not suitable for a multivariate analysis, so they were eliminated.

Evaluation Metric

The Root Mean Square Error metric (or RMSE) is used to compare and evaluate the models' performance on a 20-day prediction window. The equation for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

In this use case, **n** represents the total data points, $\mathbf{y_j}$ is the actual ticker price, and $\hat{y}$ is the predicted price for day **j**.

RESULTS

Sentiment Analysis

To conduct a comprehensive analysis of Reddit and financial news data, two evaluation methods—accuracy and F1 score—were implemented to assess and select the optimal sentiment analysis model. We used VADER for Reddit data and FinBERT for news data. Additionally, we performed manual labeling to compare the accuracy of automated labeling with that of a pretrained model, providing an overall perspective on the effectiveness of each approach.

Sentiment Analysis on Reddit Data

The process began with cleaning the Reddit texts by removing URLs, emoticons, symbols, and stopwords. VADER was then applied to obtain a composite sentiment score. A representative sample of the dataset was manually labeled with true sentiments (*true_sentiment*), allowing us to compare the model's predictions with the true labels to calculate precision, recall, and F1-Score metrics.

Sentiment	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.88	0.70	0.78	10
Negative	0.76	0.88	0.81	32
Neutral	0.84	0.73	0.78	22
Total	0.80	0.80	0.80	64

***Table 1**. Classification Report for Reddit Data*

The results show a high precision of the VADER model in classifying positive, negative, and neutral sentiments in Reddit texts. Particularly, the model performed excellently in classifying negative sentiments, which is crucial for financial analysis.

Sentiment Analysis on News Data

For the news data, FinBERT was applied after cleaning the texts to remove unwanted elements. A representative sample of the news articles was manually labeled, allowing us to compare the model's predictions with these labels to calculate performance metrics.

Sentiment	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.45	0.47	0.46	19
Negative	0.00	0.00	0.00	27
Neutral	0.68	0.92	0.78	73
Total	0.64	0.64	0.64	119

***Table 2.** Classification Report for News Data*

While the FinBERT model showed decent overall accuracy in classifying neutral sentiments, its performance in classifying positive and negative sentiments was lower. This suggests the need for further tuning of the model to improve its predictive capacity in these areas.

Pretrained Model

An additional pretrained model was used to calculate accuracy through an automated approach. The pretrained model used was 'distilbert-base-uncased-finetuned-sst-2-english' from Hugging Face, known for its sentiment analysis capabilities. This model was applied to both Reddit and news data.

Pretrained Model Results

Method	Reddit Data	News Data
Manual Labeling	0.80	0.64
Pretrained Models	0.25	0.35

Table 3. *Accuracy Comparison between Manual Labeling and Pretrained Models*

Interpretation of Results

The comparison shows that manual labeling outperforms pretrained models in both datasets, suggesting manual labeling's reliability despite being labor-intensive. Pretrained models, while convenient, displayed lower accuracy, particularly for Reddit data, highlighting the need for fine-tuning to better capture data-specific nuances.

Overall, FinBERT's performance on financial news data was superior to VADER's performance on Reddit data, underscoring the importance of domain-specific sentiment analysis. These findings suggest that while automated methods offer efficiency, manual processes and model fine-tuning are crucial for accurate sentiment evaluation in specialized contexts.

The evaluation results highlighted that, while the models demonstrated promising performance, there are areas for improvement, especially in classifying negative and positive sentiments in the news data. Below are the confusion matrices for the Reddit and news data, which offer a clear visualization of the models' performance. The colors in the confusion matrices indicate the number of correctly classified examples (on the diagonal) and misclassifications (off the diagonal). For the Reddit data, the VADER model demonstrates a high rate of correct classification for negative sentiments. Conversely, for the news data, the FinBERT model encounters challenges in accurately classifying both positive and negative sentiments.

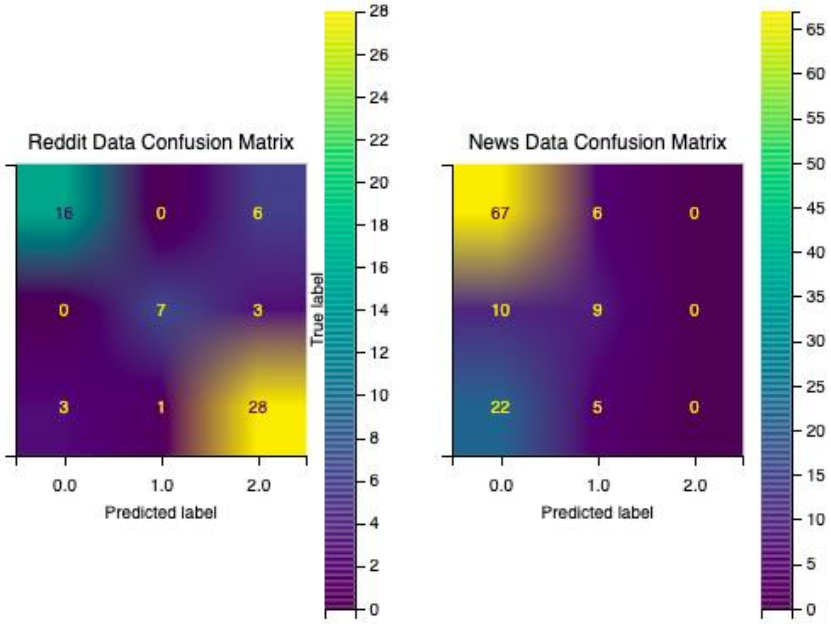


Figure 4. *Confusion Matrix for Manually labeled data*

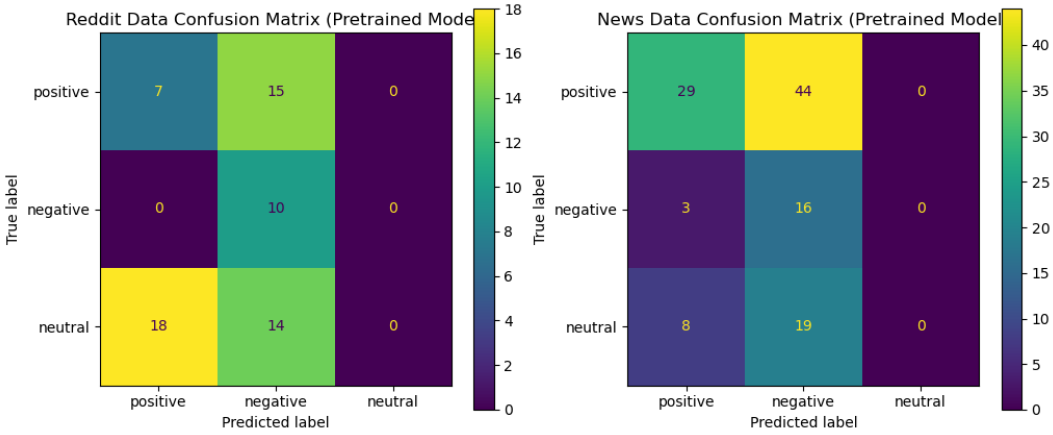


Figure 5. *Confusion Matrices for Pretrained Models*

Furthermore, the pie charts reveal that Reddit has a larger share of neutral posts compared to news headlines. This is surprising, given that social media platforms are perceived as more polarizing. While this assumption holds true for subreddits like r/wallstreetbets, extending the research to other, more neutral subreddits reveals that posts tend to be more informative. Another explanation is that this sentiment analysis was based solely on Reddit post titles, many of which contain only a few words. As a result, much of the sentiment may be lost without access to the full content and comments. This limitation will be discussed further in the research.

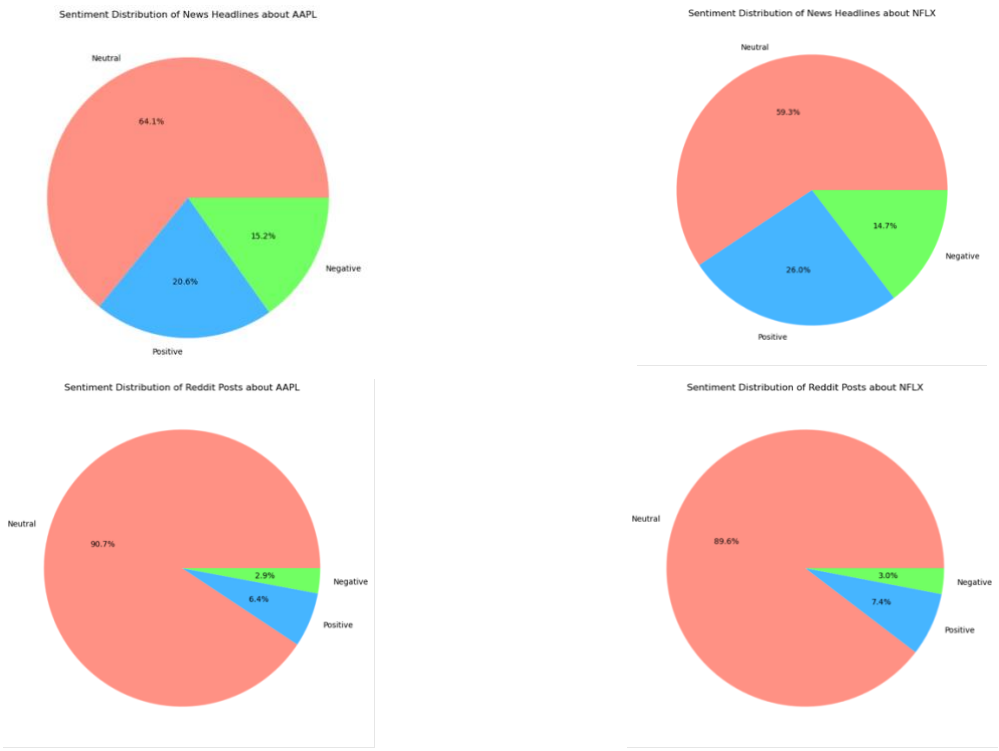


Figure 6. *Sentiment Distributions of news headlines and reddit posts*

Time Series Analysis

Exploratory Data Analysis

Utilizing the sentiment scores derived from Reddit posts and financial news articles, along with the adjusted closing prices of stocks for each day, the data was fitted and trained using a variety of machine learning models. The aim was to compare the effectiveness of these models and to determine the most effective algorithm for stock price movement prediction. The data was split into training and testing sets using a stratified sampling approach to ensure that the distribution of classes was maintained.

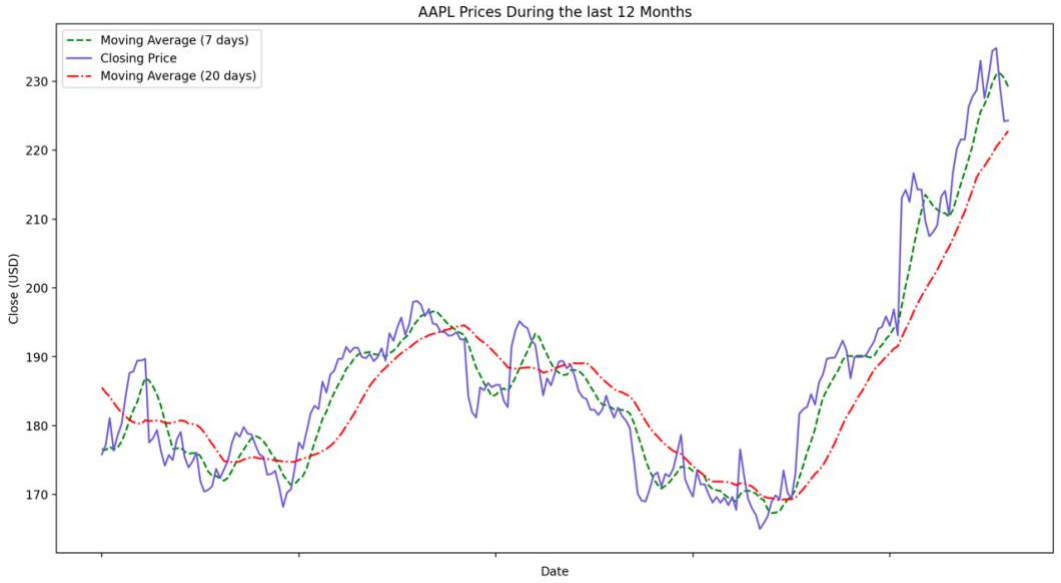


Figure 7. *Apple (AAPL) prices from July 2023 to July 2024*

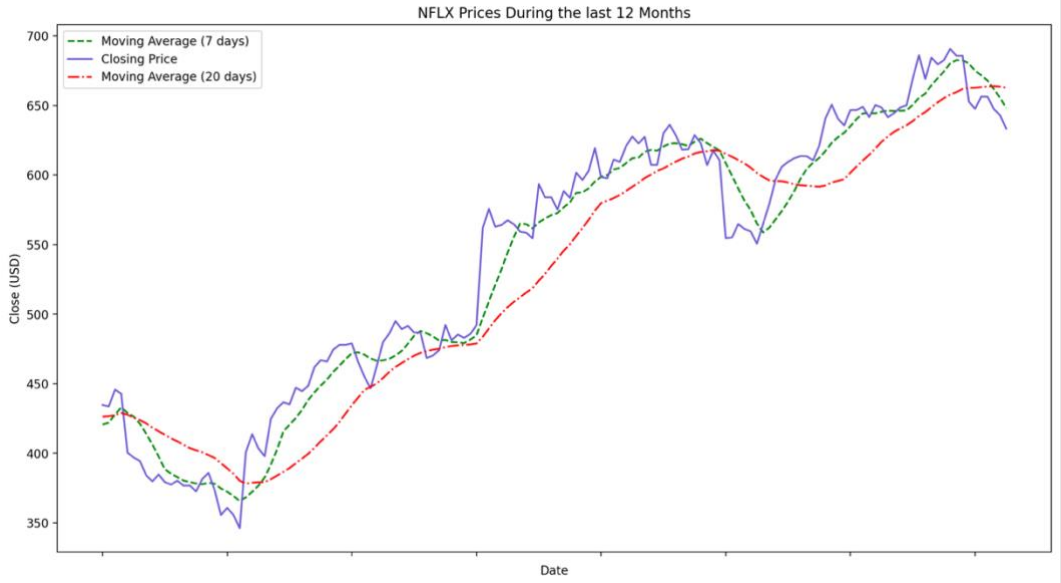


Figure 8. *Netflix (NFLX) prices from July 2023 to July 2024*

To enhance the visualization, we included additional metrics – the 7-day and 20-day moving averages – to the plot. These moving averages smooth the price data and provide a clearer depiction of stock trends. These two technical indicators and other similar metrics were also evaluated during the modeling process but were eventually excluded due to high correlation with the existing price features.

Over the last 12 months, AAPL's stock price ranged from $164.08 to $237.23, while NFLX's stock price fluctuated between $344.73 and $697.49. AAPL hit its second-lowest price in October 2023, then trended upwards until reaching its 52-week low in April 2024. The stock subsequently recovered and consistently increased in value over the last three months. In contrast, NFLX dipped to its lowest point in October 2023 and has trended upwards since then, though with more pronounced price fluctuations and volatility. These distinctive patterns align with the companies' different sectors.

The overall upward movement in both stocks indicates increased investor confidence and positive market sentiment regarding each company's performance.

In addition to the overall trend, Figure 9 shows the time series of each feature. It can be observed that 'reddit_sentiment' and 'news_sentiment' reach peaks and troughs at various times. These two series exhibit a weak correlation of 11.72% for AAPL and -7.13% for Reddit. This is an interesting observation, indicating that social media sentiment on Reddit and traditional news do not exert strong influence on each other.

Furthermore, the features 'open,' 'high,' 'low,' and 'close' show remarkably similar patterns, indicating a strong correlation between them.

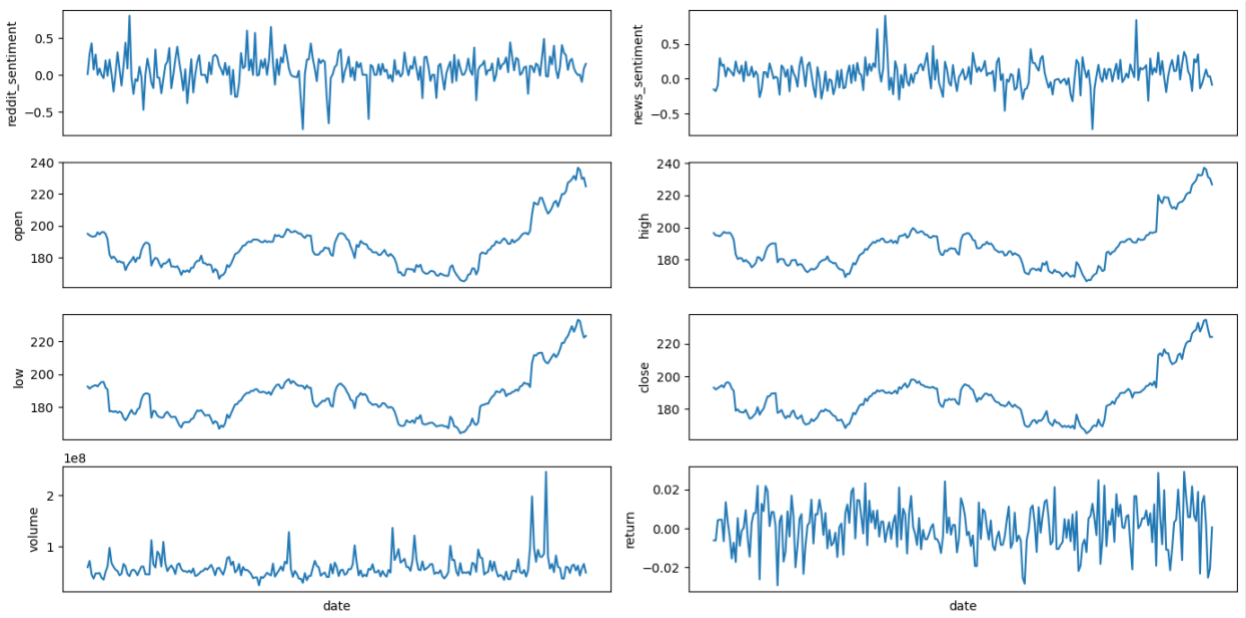


Figure 9. *Time Series of each feature*

The initial correlation matrix (Figure 10) shows highly correlated relationships between 'open' (opening price) and 'low' (lowest price) as well as 'high' (highest price), so these two features are removed.

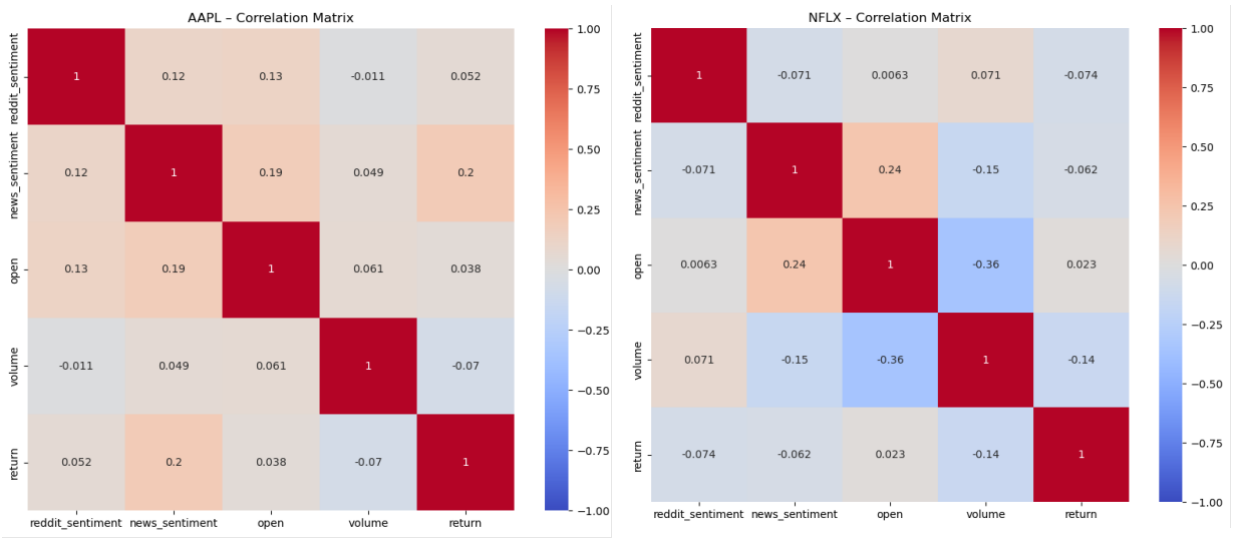
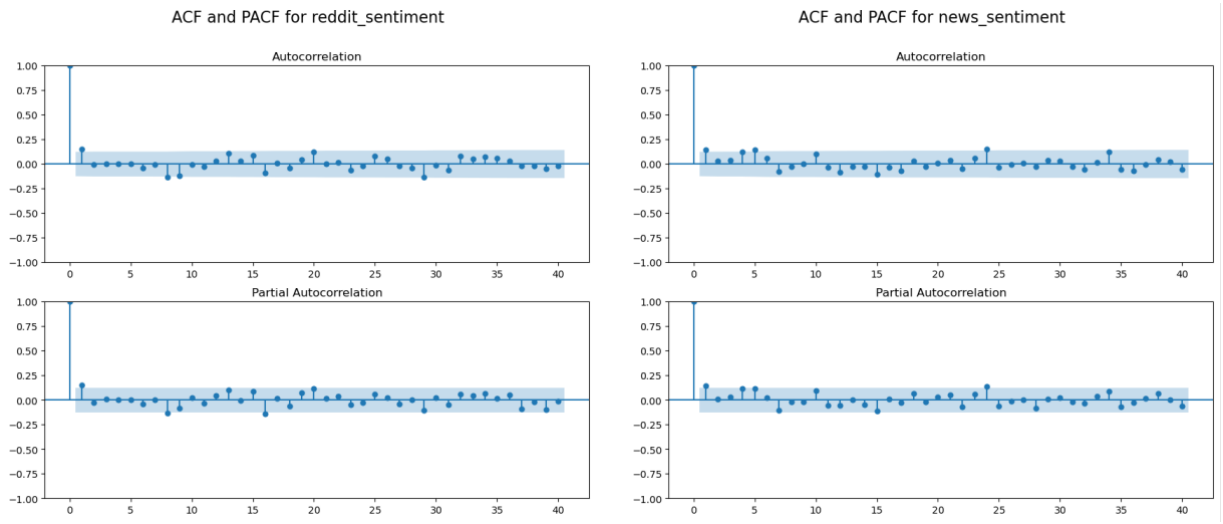


Figure 10. *Correlation Matrixes for Apple and Netflix features*

Figure 11 presents the ACF and PACF plots for the time series features to help understand the attributes of the time series. The ACF plot illustrates the correlation between a data point and its lagged values, accounting for all intermediate lags. This plot provides an overall view of how each observation is related to its past values (Brownlee, 2020). The PACF plot isolates the direct correlation between a data point and a specific lag, controlling for the influence of other lags. This plot gives a more precise picture of the relationship between a data point and its immediate past (Brownlee, 2020).



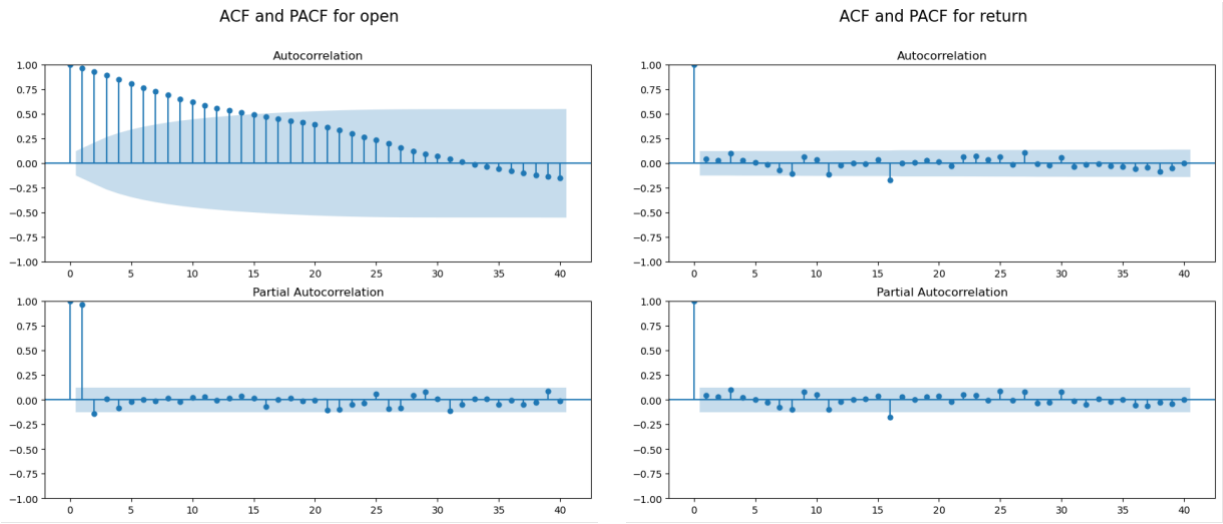


Figure 11. *ACF and PACF Plots*

These plots were utilized to examine the autocorrelation within the time series to identify trends and seasonality. For 'reddit_sentiment,' 'news_sentiment,' and 'return,' there is almost no significant non-zero correlation. The ACF and PACF plots show values remarkably close to zero for all lags, which indicate white noise. Thus, these time series are random, and they do not exhibit any predictable patterns.

However, the 'open' (opening prices) feature exhibits significant autocorrelation values indicating the strong indirect influence that past prices have on current stock prices. As expected, smaller lags have a greater impact on the current price compared to prices from a long time ago. Non-adjacent lags quickly decay to zero. Also, no seasonal patterns are evident in these plots. Therefore, transformation is not necessary since these time series look stationary.

Since stationarity is an important attribute of time series, the Dickey-Fuller test was performed as a second check. When the p-value is less than the critical value of 5%, the null hypothesis that the time series is not stationary is rejected. The results show that all the time series are stationary. After this test, there is enough confidence to conclude that the time series is stationary. Hence, there is no need to apply differencing.

The modeling phase involves applying appropriate time series models to the stationary data. The data is split into training and testing sets to evaluate model performance using RMSE on both the training and testing data. Hyperparameter tuning is performed to optimize each model to ensure that they capture the underlying patterns effectively. The results are included in Table 4.

	Apple (AAPL)		**Netflix (NFLX)**	
Prediction Models	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE
LSTM	3.3952	3.6296	15.9595	17.3492
GRU	2.9480	3.7665	14.9435	14.2629
GAN	0.2418	7.5453	3.7959	18.8177
XG Boost	0.1496	21.2603	0.4279	33.3354
Random Forest Regressor	1.2927	18.5265	6.5207	28.9260

Table 4. *Prediction models' performance results*

The RMSE scores and the corresponding plots (Figure 12) indicate that the simplest model, the Random Forest Regressor, exhibits the poorest performance. It fails to capture data trends adequately, resulting in inaccurate predictions. The other tree-based model, XGBoost, performs marginally better, but its RMSE scores remain high. The more complex neural network models—LSTM, GRU, and GAN—all show decent price prediction capabilities.

Figure 12. *Stock price prediction performance on a 20-day window*

However, both GAN and XGBoost display signs of overfitting, evidenced by near-zero RMSE scores for training but significantly higher scores for testing. The LSTM model provides the most accurate predictions for AAPL, while the GRU model outperforms others for NFLX. Overall, the GRU model demonstrates the best performance for both tickers.

CHALLENGES

Throughout this study, we encountered several challenges that required us to adapt and refine our approach. Here are the main hurdles we faced and how we overcame them:

Data Collection Limitations

Initial Data Scarcity: The data from r/wallstreetbets was insufficient to build a robust model due to the lack of consistent news flow discussing the targeted stocks, Apple, and Netflix. To gather additional data, the team decided to scrape using the 'request' and 'BeautifulSoup' libraries. This process presented several challenges, including API limitations, budget constraints, and data availability. Advanced scraping methods, such as using IP proxies, were implemented to avoid server blocks and ensure continuous data flow. Additionally, attempts to scrap data from Twitter and Stocktwits were considered but abandoned due to high cost and lack of access. Despite these efforts, many approaches failed, and some previously accessible libraries were no longer functional.

To address this issue, the data collection was expanded to include other popular investing subreddits such as r/stocks, r/investing, r/Daytrading, and r/StockMarket. This expansion helped compile a more comprehensive dataset, resulting in 1,945 posts about Apple and 434 posts about Netflix.

Data Quality

Emoticons and Sentiment Analysis: Emoticons are commonly used on social media to express feelings and tone. Initially, these were treated as noise and removed during data preprocessing. However, it was later realized that emoticons carry valuable sentimental information. By including and properly processing them, it became possible to capture more nuanced expressions of sentiment in the data.

URLs and Stopwords: URLs and stopwords initially disrupted our sentiment analysis accuracy. To address this issue, extensive data cleaning was undertaken to remove these elements. This process improved the quality of our input data and enhanced the reliability of our sentiment scores.

Methodological Adjustments

From Classification to Regression: Our initial plan was to classify stock purchase decisions based on sentiment analysis. However, we faced challenges with data suitability and alignment with the study's objectives. Consequently, we shifted to a time series regression approach. This new method involved predicting stock prices based on sentiment scores from Reddit posts and news articles, allowing for a more continuous and comprehensive analysis of market behavior.

Model Performance

Pretrained Sentiment Models: Using pretrained sentiment models was convenient, but initially, they did not provide the necessary accuracy. These models struggled with financial jargon and nuanced expressions

specific to our context. Consequently, we had to continuously evaluate and fine-tune these models to improve their performance for our application.

Manual Labeling and Validation

Labor-Intensive Manual Labeling: Manually labeling data was time-consuming and prone to human error. Despite the effort required, this step was crucial for validating our sentiment analysis models. Ensuring consistency and objectivity in manual labeling demanded meticulous attention to detail and repeated refinements.

Data Integration

Combining Multiple Data Sources: Integrating sentiment scores from various sources was challenging. The expression of sentiment in news articles differs significantly from that in social media posts, necessitating careful normalization and calibration to ensure the aggregated sentiment scores were meaningful and reliable. By addressing these challenges with thoughtful adjustments and improvements in our data collection and analysis processes, we were able to develop a robust and reliable predictive model for stock prices based on sentiment analysis from multiple influential sources.

CONCLUSION

The study demonstrates that manually labeling data provides higher accuracy in sentiment analysis results compared to using pretrained models. The accuracy of manual labeling for Reddit data was 0.80 and for news data was 0.64. In contrast, pretrained models showed an accuracy of 0.25 for Reddit data and 0.35 for news data. Although manual labeling is more precise, it is labor-intensive and time-consuming. Pretrained models are more efficient in terms of time but require additional tuning to improve their accuracy.

Gated Recurrent Unit (GRU) proves to be the best-performing model for stock price prediction and multivariate time series analysis even with a small dataset. It effectively captures underlying trends and demonstrates robust prediction accuracy without overfitting, making it the optimal choice for forecasting stock prices.

REFERENCES

- Brownlee, J. (2020, August 14). *A gentle introduction to autocorrelation and partial autocorrelation*. Retrieved from https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/
- Fan, W., & Gordon, M. D. (2014). The power of social media analytics. Communications of the ACM, 57(6), 74-81.
- García-Méndez, M., De Arriba Pérez, F., & González-Castaño, F. J. (2022). Media bias and its influence on stock markets: Evidence from the COVID-19 pandemic. Journal of Financial Markets, 56, 100619.
- Hasso, T., Müller, D., Pelster, M., & Warkulat, J. (2022). Who let the bulls out? Bitcoin speculation on r/wallstreetbets. Finance Research Letters, 38, 101409.
- Hu, Y., Tripathi, S., Ding, Y., & Gan, Q. (2021). Influence of social media on stock prices: The case of Netflix. Journal of Financial Economics, 140(3), 673-692.
- Li, J., & Pan, W. (2022). The Impact of Financial News on Stock Prices: Evidence from the COVID-19 Pandemic. Journal of Financial Markets, 56, 100619.
- Li, Liang & Hu, Changming & Hou, Yajun. (2020). Prediction analysis of shield vertical attitude based on GRU. Journal of Physics: Conference Series. 1651. 012032. 10.1088/1742-6596/1651/1/012032.
- Ma, Z., Sun, Q., & Sun, L. (2020). The impact of social media on stock prices: A case study of Apple. Finance Research Letters, 34, 101282.
- Mittal, A., Goel, S., & Katal, D. (2021). Predicting Stock Market Trends Using Social Media Analysis. International Journal of Information Management, 58, 102311.
- Quiver Quantitative (2022). The WallStreetBets Quantitative Strategy https://seekingalpha.com/article/4521071-wallstreetbets-quantitative-strategy.
- Vraj Sheth, Urvashi Tripathi, Ankit Sharma. (2022) A Comparative Analysis of Machine Learning Algorithms for Classification Purpose, Procedia Computer Science, Volume 215, 2022, Pages 422-431, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.044.
- Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Retrieved from https://vadersentiment.readthedocs.io/en/latest/.
- Olah, C. (2015, August 27). *Understanding LSTM networks*. Understanding LSTM Networks . https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- Yang, X., Si, L., Yu, H., & Zhao, L. (2020). FinBERT: A Pretrained Language Model for Financial Communications. Retrieved from https://huggingface.co/ProsusAI/finbert.
- Sonkiya, P., Bajpai, V., & Bansal, A. (2021, July 18). Stock price prediction using Bert and gan. https://arxiv.org/pdf/2107.09055.pdf