TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF EQUATIONS

INTRODUCTION

Stock Market Overview

The stock market plays a crucial role in global finance. It indicates economic health and is used as mechanism for companies to raise capital. Fundamentally, the stock market facilitates the buying and selling equities or stocks. Stocks represent the fractional ownership in companies. Shares entitle the holder to a proportional stake in the company's assets and profits corresponding to the amount of stock owned. The stock price reflects the market's current value of a share, which fluctuates based on various factors such as corporate performance, macroeconomic trends, and investor sentiment. (Bodie, 2014). Prices are determined by stock exchanges, where buyers and sellers interact. A well-functioning stock market is seen as a critical indicator of a nation's economic health, as it reflects the collective sentiment of investors and their expectations for future economic conditions (Bodie, 2014). Additionally, stock markets play a crucial role in economic development since it provides companies with access to capital.

Both financial metrics and external factors influence the stock market dynamics. The law of supply and demand is the key principle that determine stock prices. When the demand for a stock exceeds its supply, the price increases. Conversely, when supply exceeds demand, prices decline (Graton, 2022). Investor sentiment towards a company's performance is a significant determinant of the demand. For instance, positive financial reports increase investor confidence, and it leads to raise demand and stock prices. On the other hand, negative news can lead to reduce demand and stock prices.

In addition to traditional financial metrics, investor psychology plays a significant role in stock market behavior. Behavioral finance proposes that irrational factors, such as optimism, fear, and herd behavior, can move stock prices away from their intrinsic value (Akin 2024). This deviation is especially observable during economic and speculative bubbles, where prices may rise far above their actual value. They crash when the bubble bursts. A historical example of this phenomenon is the Dotcom Bubble, where technology stocks experienced significant overvaluation due to excessive speculation, leading to a market crash in the early 2000s (Delong 2006). Such events highlight the influence of behavioral factors on market volatility.

Stock Returns

Stock returns represent the gain or loss on an investment over a specific period. They are mostly used for evaluating portfolio performance. Returns arise from two primary sources: capital gains, where investors profit from the increase in stock price, and dividends, which are periodic payments distributed by companies to shareholders (Hayes, 2023). The formula for calculating total stock returns is:

Equation 1: Total Stock Return

$$Total\ Stock\ Return = \frac{(Ending\ Price - Initial\ Price) + Dividends}{Initial\ Price} \times 100$$

Capital appreciation represents the change in stock price, while dividends provide periodic income to shareholders. Investors evaluate total stock returns to assess their performance relative to the market. This table shows the capital gains and total dividends paid from January 1, 2015, to December 31, 2020, for the selected companies in this study.

Company	Capital Gain (%)	Total Dividends Paid
Apple	**305.48**	**3.95**
Amazon	**438.42**	**0.0**
Tesla	**1602.91**	**0.0**
Microsoft	**373.78**	**10.5**
Google	**143.14**	**0.0**
Netflix	**851.67**	**0.0**

Figure 1: Capital Gains and Total Dividends Paid From January 1 2015 to December 31 2020 for the Selected Companies

The Role of Social Media Sentiment in Stock Markets

In recent years, there has been a growing interest in non-traditional data sources, mainly social media sentiment, and its influence on stock market dynamics. The rise of platforms such as Twitter has enabled the rapid circulation of public sentiment, leading to immediate effects across financial markets (Ranco et al., 2015). Both investors and market speculators utilize social media to share opinions, forecasts, and reactions about

macroeconomic events and company-specific news. This real-time sentiment data, which are frequently shared before formal news or financial reports, can act as a leading indicator of stock price movements (Bollen, Johan et al., 2011).

Research indicated that social media platforms play a crucial role in shaping investor sentiment. Sentiment analysis of Twitter data has consistently shown a correlation with market behavior, especially regarding short-term stock price fluctuations. As a result, both investors and traders are increasingly leveraging sentiment metrics to determine their trading strategies. They rely on real-time public opinion to predict market trends and make investment choices.

Social Media and Finance: The Role of Alternative Data

The emergence of alternative data has changed the way of financial analysis in recent years. Traditionally, investors have relied on company-sourced data, such as earnings reports, financial statements, and regulatory documents in order to make informed decisions. However, in the digital era, non-traditional data sources, such as social media, web traffic, and consumer reviews, have become invaluable for predicting market movements (Gosset, 2023). Among these, social media has a special place for its accessibility, instancy, and widespread use.

Platforms like Twitter provide vast storage of real-time sentiment data which allow investors to observe the mood of the market. Twitter has become an important tool for investors to observe public opinion and sentiment about individual stocks, industries, and economic trends. The platform enables users to quickly share opinions on market events, corporate news, and macroeconomic developments, as fast as in real time. They can influence trading decisions instantaneously.

Recent research has confirmed the importance of public sentiment from social media in predicting stock price movements. For instance, studies have shown that positive sentiment on Twitter often correlates with upward movements in stock prices, while negative sentiment frequently leads to declines (Zhang et al., 2021). The timeliness and accessibility of these sentiments provide a competitive advantage for traders. As a result, sentiment analysis derived from platforms like Twitter is increasingly used alongside traditional data sources to create more comprehensive models for predicting stock returns (Ruiz et al., 2012).

Traditional economic indicators, such as quarterly earnings reports or GDP statistics, often suffer from time lags. Whereas social media provides real-time insights for market sentiment. This allows traders to anticipate market shifts and respond more rapidly, offering a competitive edge in fast-moving financial environments (Tetlock, 2007). The current study aims to contribute this field by analyzing how well Twitter sentiment predicts stock returns for some of the most traded companies on the NASDAQ.

Research Motivation

Several high-profile figures have illustrated the potential impact of social media on stock prices. For example, tweets from influential figures, such as Elon Musk, have caused significant fluctuations in Tesla's stock price. It is a good example of the power of social media in influencing market behavior (Metta et al., 2022). However, while Musk's tweets were such an interesting case, this research expands its focus beyond individual influencers. Instead, it evaluates broader public sentiment regarding to multiple companies. It aims to test the hypothesis that aggregate sentiment on Twitter can predict stock price changes across different companies, not just those influenced by highly influential figures.

Understanding how public sentiment shapes stock returns could be invaluable for participants in the market such as investors and financial analysts. For these groups, leveraging alternative data sources such as social media sentiment could offer significant advantages in trading strategies or better understanding the market. Additionally, this research contributes to the growing body of literature that examines the intersection of social media and financial markets. It aims to provide insights about how non-traditional data sources can support established forms of market analysis.

Research Objectives

The main objective of this research is to find out whether public sentiment on Twitter can predict stock returns for six leading companies: Apple, Amazon, Google, Microsoft, Tesla, and Netflix. The following are the goal this study seeks to accomplish:

1. Classify Public Sentiment: This study will use machine learning models, specifically RoBERTa, to categorize tweets about the six companies into three groups: positive, neutral, and negative.

2. Correlate Sentiment with Stock Price Movements: The research will determine whether shifts in the stock price are correlated with changes in the public's sentiment. Statistical methods, such as Granger Causality Tests, which will be used to examine the predictive relationship between sentiment and stock prices.
3. Evaluate the Predictive Power of Sentiment Data: The research will evaluate if Twitter sentiment can serve as a leading indicator for stock performance.

Both sentiment analysis and financial time-series data will be utilized to test the hypothesis that aggregate public sentiment on social media can offer predictive insights into stock returns. This flowchart visually represents the research methodology:

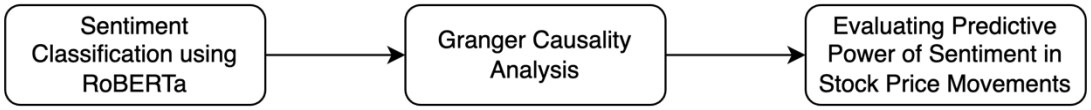

Figure 2: Flowchart for the Research Methodology

LITERATURE REVIEW

The growing availability of alternative data and advancements in data processing have transformed financial market analysis. Recent literature demonstrates how social media sentiment can be utilized to predict financial market outcomes with researchers developing new models and methodologies to take advantage of this data.

Kalampokis, Tambouris, and Tarabanis (2013), in their paper “Understanding the Predictive Power of Social Media,” explored how social media can predict various real-world outcomes, including stock market volatility. While acknowledging the predictive potential of social media, they emphasized that accurate predictions require advanced sentiment analysis techniques and refined data filtering. Their findings emphasize the need for strong methodologies to manage the noisy, informal nature of social media platforms, which can prevent precise sentiment extraction. Improvements in data conditioning and sentiment classification are essential to leverage social media for prediction effectively.

Building on these foundations, Mittal and Goel (2017) explored sentiment analysis for predicting movements in the Dow Jones Industrial Average (DJIA). They utilized a

Self-Organizing Fuzzy Neural Network (SOFNN) to examine Twitter sentiment and its influence on stock market behavior. The SOFNN model achieved 75.56% accuracy in predicting the DJIA's directional movements. Their research emphasized the role of natural language processing (NLP) in financial forecasting and demonstrated the value of machine learning models in utilizing social media data to enhance market predictions. This work further illustrates the growing importance of integrating sentiment analysis with machine learning for financial forecasting.

Similarly, Guo, Sun, and Qian (2017) applied the Thermal Optimal Path (TOP) method to explore how investor sentiment impacts stock prices in China's stock market. Using sentiment data from Xueqiu, a popular Chinese social media platform, they showed that sentiment's predictive power fluctuates with investor attention and market conditions. Their findings suggest that sentiment is a dynamic predictor, varying in effectiveness based on market interest and external factors, such as economic conditions. These results indicate that sentiment-driven stock predictions are sensitive to the levels of public attention, emphasizing the complexity of using sentiment data for financial forecasting.

Bhadkamkar and Bhattacharya (2022) focused on individual stocks, specifically Tesla, to study the influence of Elon Musk's tweets on stock price movements. Their time-series analysis revealed a direct correlation between sentiment shifts, which are driven primarily by Musk's tweets, and subsequent changes in Tesla's stock value. The study highlighted the impact of the high-profile figures can have on stock dynamics, illustrating the importance of source-based sentiment analysis. In contrast, Edman and Weishaupt (2020) found that Twitter sentiment was an unreliable predictor of Tesla's within a single day of stock returns. Their study used Lasso regression and Granger causality tests, suggesting that the predictive power of sentiment can vary depending on the methodology and the scope of the analysis.

As deep learning models such as BERT and RoBERTa have gained recognition, new possibilities for stock market forecasting have emerged. These transformer-based models have shown effectiveness in natural language processing tasks, particularly in understanding and analyzing sentiment from short-form social media content like tweets (Devlin et al., 2019). RoBERTa, specifically, has been proven to capture sentiment more accurately, making it highly suitable for real-time financial analysis. The use of RoBERTa

in this study enhances the ability to process and interpret the sentiment from Twitter data, which is crucial for understanding how public opinion affects stock prices.

Studies applying deep learning models in financial forecasting have demonstrated promising results, with such models identifying complex patterns within large datasets. Research shows that these models can manage the volatility of financial markets, making them indispensable tools for traders and researchers (Liu et al., 2020). Transformer models, particularly RoBERTa, stand out for their ability to process informal, abbreviated social media language, offering an advantage over traditional machine learning techniques. Their flexibility and adaptability make them well-suited for highly dynamic financial environments, where rapid shifts in public sentiment can influence market performance.

Incorporating alternative data sources has become a key focus in financial markets. J.P. Morgan's *Alternative Data Handbook* (2019) highlights the importance of utilizing data from unconventional sources, such as social media activity, satellite imagery, and credit card transactions, to inform investment strategies. These alternative data streams offer unique, real-time insights into consumer behavior, economic trends, and market sentiment, providing a competitive edge. J.P. Morgan emphasizes the need for sophisticated analytical techniques to filter and translate this raw data into actionable investment decisions, emphasizing the transformative potential of alternative data in modern financial analysis.

Compared to the reviewed literature, this thesis offers several distinctions. First, unlike Bhadkamkar's focus on Elon Musk's tweets, this study analyzes sentiment from general Twitter users. Additionally, rather than examining stock indices like Mittal and Goel's focus on the DJIA, this research looks individually at six of the most popular tech stocks traded on the NASDAQ. This paper also employs bidirectional Granger causality to assess the predictability of sentiment and stock returns, in contrast to Guo, Sun, and Qian's use of the Thermal Optimal Path method. Furthermore, this research moves beyond simple sentiment clustering by eliminating neutral tweets and constructing two sentiment differential signals, one weighted and one unweighted by retweets. This approach allows for a more subtle analysis of sentiment and its impact on stock market performance. By leveraging modern machine learning techniques and alternative data

sources, this thesis seeks to contribute to the ongoing research about the predictive power of social media sentiment in financial markets.

DATASET

The data collection process for this study involved two primary components: the extraction of stock price and transaction volume data and the gathering and preprocessing of tweets related to the six target companies - Apple, Amazon, Google, Microsoft, Tesla, and Netflix. These datasets form the basis for the sentiment analysis and predictive modeling which are necessary to explore the relationship between public sentiment and stock price movements.

Stock Price Data

Historical stock data for each of the six companies was retrieved using the 'yfinance' Python library, which is a widely used tool that accesses historical market data from Yahoo Finance. This enabled the collection of detailed stock information for the period covering from January 1, 2015, to December 31, 2020. It is consistent with the sentiment data derived from Twitter over the same timeframe.

The data retrieved was structured into pandas Data Frames for each company, containing the following key columns:

- Date: The trading day timestamp.
- Open Price: The stock price at the start of the trading session.
- Close Price: The price at the end of the trading session.
- High Price: The highest price reached during the session.
- Low Price: The lowest price reached during the session.
- Volume: The total number of shares traded on that day, representing both buy and sell transactions.
- Adjusted Close Price: The closing price adjusted for corporate actions, such as stock splits and dividend payments. It provides a more accurate long-term view of stock value.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2015-01-02	15.629	15.738	15.348	15.426	15.426	55664000
1	2015-01-05	15.351	15.419	15.042	15.109	15.109	55484000
2	2015-01-06	15.112	15.15	14.619	14.764	14.764	70380000
3	2015-01-07	14.875	15.064	14.767	14.921	14.921	52806000
4	2015-01-08	15.016	15.157	14.806	15.023	15.023	61768000

Figure 3: Stock Price Data for Amazon rounded to 3 significant figure

For the purposes of this study, the focus was narrowed to three specific columns: Date, Adjusted Close Price, and Volume. These variables are critical for performing time series analysis, which aims to assess stock price movements in relation to sentiment data. The adjusted close price was prioritized over the raw closing price due to its incorporation of factors like stock splits and dividends, offering a more accurate reflection of value changes over time (Bodie et al., 2014).

Additionally, the volume data was essential for assessing market liquidity and investor interest. Large fluctuations in trading volume are often correlated with significant price changes, making it an important metric for understanding the context in which sentiment shifts may influence stock returns (Alvarez et al., 2023).

Tweet Data Collection

The second major component of the dataset involved the collection and preparation of relevant tweets related to each of the six companies. For this, a publicly available dataset from Kaggle, titled ‘Tweets about the Top Companies from 2015 to 2020’, was utilized. This dataset contains millions of tweets mentioning prominent companies, offering a rich source of textual data that captures public sentiment towards the selected stocks. The timeframe of the tweets aligns with the stock price data, covering the years from 2015 to 2020.

Tweet Filtering and Preprocessing

To ensure that only relevant tweets were included in the analysis, a Boolean mask was applied to filter tweets based on company mentions. The text of each tweet was

scanned for the appearance of the company's ticker symbol (e.g., $AAPL for Apple, $AMZN for Amazon). This process was repeated for each of the six target companies:

- Apple: $AAPL - Amazon: $AMZN - Google: $GOOG
- Microsoft: $MSFT - Tesla: TSLA - Netflix: $NFLX

This Boolean mask filtered the dataset, ensuring that only tweets relevant to each company were included in the subsequent analysis. Once filtered, the relevant tweets were stored in a pandas DataFrame for each company, with the following fields:

- Date: The timestamp indicating when the tweet was posted.
- Text: The content of the tweet, which would later be subjected to sentiment analysis.
- Engagement Metrics: This includes the number of retweets, likes, and replies each tweet received, serving as indicators of how much attention the tweet gathered within the broader Twitter community.

	post_date	text	comment_num	retweet_num	like_num	ticker_symbol	date	apple	amazon	tesla	microsoft	google	None	netflix
0	2015-01-01 00:00:57	lx21 made $10,008 on AAPL -Check it out! htt...	0	0	1	AAPL	2015-01-01	True	False	False	False	False	False	False
1	2015-01-01 00:01:36	Insanity of today weirdo massive selling. $aap...	0	0	0	AAPL	2015-01-01	True	False	False	False	False	False	False
2	2015-01-01 00:01:50	S&P100 #Stocks Performance $HD LOW $SBUX TGT...	0	0	0	AMZN	2015-01-01	False	True	False	False	False	False	False
3	2015-01-01 00:06:47	$GM TSLA: Volkswagen Pushes 2014 Record Recal...	0	0	1	TSLA	2015-01-01	False	False	True	False	False	False	False
4	2015-01-01 00:10:05	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1	AAPL	2015-01-01	True	False	True	False	False	False	False

Figure 4: Twitter Data

Data Aggregation and Weighting

The engagement metrics played a key role in the subsequent sentiment analysis. Tweets with higher engagement, such as a large number of retweets, are more likely to represent broader public sentiment. They may have a greater impact on investor perceptions and, potentially, stock price movements (Bollen, Johan et al., 2011). Therefore, particularly the number of retweets was used to weight the importance of certain tweets during the sentiment analysis phase. This weighting ensured that tweets which gained substantial attention on Twitter were given more influence in the model's sentiment classification.

Challenges in Data Preprocessing

During the preprocessing phase, several challenges were encountered both in the stock price data and the tweet data. One of the key challenges was the presence of noise in the tweet dataset. Twitter often contains irrelevant or spam content, and some tweets mentioning stock ticker symbols may not directly relate to stock market discussions. To fix this issue, tweets filtered with more than 30 retweets or 50 likes to focus on high-impact tweets.

Another significant challenge involved missing data in the stock price dataset. On certain trading days, especially around holidays or market closures, data for stock prices was not available. To address this, spline interpolation was applied to estimate missing values, ensuring a continuous time series for analysis (Bailey, 2022). This approach ensured that no gaps in the data would affect the time series analysis or predictive modeling.

Finally, the sentiment analysis itself faced challenges due to ambiguous language within the tweets. Twitter often has informal language, sarcasm, or irony, which can complicate sentiment classification. Even though the RoBERTa model used for sentiment analysis is trained to handle social media language, certain variations in tweets may still lead to misclassification. Further refinement of the sentiment analysis, such as the use of manual labeling or more advanced context-aware models, could improve classification accuracy in future studies.

Sentiment Analysis Models: RoBERTa Model

The RoBERTa (Robustly Optimized BERT Approach) model marks a significant enhancement over the original BERT (Bidirectional Encoder Representations from Transformers) model, introduced by Google in 2018 (Devlin et al., 2018). BERT revolutionized the field of natural language processing (NLP) by allowing bidirectional text processing. Researchers identified several areas for improvement especially regarding the training process. This led to the development of RoBERTa, which integrates a series of optimizations to boost the model's performance. These make it especially suited for social media content like tweets.

RoBERTa retains BERT's foundational model, which is built upon the Transformer model. This architecture leverages self-attention mechanisms. These allow RoBERTa to assign varying importance to words based on their connections to other words. It doesn't matter where the words appear in a sentence. This feature is useful for understanding the context of words in informal or shortened social media language.

Key Improvements in RoBERTa Over BERT

1. Larger and More Diverse Training Data: One of the primary advancements RoBERTa introduces is the use of a significantly larger and more diverse training dataset compared to BERT. BERT's training data primarily consisted English Wikipedia and BookCorpus (Jha, 2024). RoBERTa incorporated additional sources such as Common Crawl, OpenWebText, and other web-based texts (Grigory 2019). This broader range of data exposure enhances RoBERTa's ability to understand various text styles, domains, and even informal language which makes it especially relevant for analyzing social media platforms like Twitter (Liu et al., 2019).
2. Dynamic Masking: RoBERTa implements dynamic masking in the Masked Language Modeling (MLM) task, where different words are masked each time the model encounters the same sentence during training (Grigory 2019). In contrast, BERT uses static masking, where the same words are always masked in repeated sentences (Jha, 2024). This dynamic approach in RoBERTa improves the model's robustness and equips it to generalize better across different contexts. The enhancement is useful in processing fluid, context-dependent social media text (Liu et al., 2019).
3. Longer Training Durations and Larger Batch Sizes: Another improvement is the extension of training time and an increase in batch sizes. These adjustments enable RoBERTa to capture more intricate language patterns and better understand complex word relationships (Grigory 2019). This is important in sentiment analysis, where subtle differences in wording can alter the sentiment of a tweet. For example, a small variation between "This stock is rising fast!" and "This stock is rising fast?" can convey vastly different sentiments, and RoBERTa's extended training allows it to detect such variations.

4. Exclusion of Next Sentence Prediction (NSP): In BERT's training, Next Sentence Prediction (NSP) was used alongside MLM (Jha, 2024). However, research indicated that NSP contributed minimally to performance improvements. Therefore, RoBERTa omits NSP entirely, focusing solely on MLM, which has been shown to enhance language modeling capabilities (Liu et al., 2019). This allows the model to allocate more training resources to improve its understanding of masked words, further enhance its accuracy for context understanding.

Fine-Tuning RoBERTa for Sentiment Analysis

RoBERTa's power for sentiment analysis lies in its ability to be fine-tuned on specific datasets, adapting its pre-trained knowledge to a specialized task such as sentiment classification. In this research, RoBERTa is fine-tuned on a labeled dataset of tweets concerning six major companies - Apple, Amazon, Google, Microsoft, Tesla, and Netflix. The goal of the fine-tuning process is to classify the sentiment of each tweet as positive, negative, or neutral.

For this study, the fine-tuning was conducted using Python's 'Hugging Face' library, a popular toolset for working with pre-trained models like RoBERTa. Hugging Face provides a user-friendly interface for applying models to custom datasets, making it ideal for this type of sentiment analysis. By classifying tweets related to the six companies, the study aims to explore whether public sentiment on Twitter can serve as an indicator of stock market performance.

METHODOLOGY:

Applying RoBERTa for Sentiment Classification

The pre-trained RoBERTa model from Hugging Face's Transformer library was used to classify the sentiment of tweets related to six target companies - Apple, Amazon, Google, Microsoft, Tesla, and Netflix. The model, already fine-tuned on 124 million tweets, was suitable for social media analysis, especially considering the informal, abbreviated, and ambiguous nature of many tweets (Liu et al., 2019). The goal was to categorize each tweet as positive, negative, or neutral to facilitate the analysis of how public sentiment impacted stock performance.

1. Initialization of the Model and Tokenizer

The first step was initializing the pre-trained RoBERTa model and its associated tokenizer. Tokenization is essential in converting raw text into numerical tokens that the model can interpret. Given the informal nature of Twitter, this tokenizer was particularly useful for handling abbreviations, slang, and inconsistent grammar, ensuring that the sentiment analysis accurately captured the nuances in the tweets (Howard & Ruder, 2018).

2. Tokenization

The tokenization process involved breaking down the text into manageable units, tokens, which represented either whole words or subword units. These tokens were mapped to their respective indices within RoBERTa's pre-defined vocabulary. Additionally, the following steps were included:

- Packing sequences to ensure consistency in input length for batch processing.
- Creating attention masks to differentiate meaningful tokens from packing elements, ensuring the model focuses on the relevant parts of each tweet.

This preparation allowed RoBERTa to interpret the tweets contextually and accurately, providing the foundation for sentiment classification.

3. Model Inference

Following tokenization, the tokenized tweets were fed into the RoBERTa model to generate sentiment predictions. The model outputted logits –'unnormalized scores ', representing the likelihood of each sentiment category (positive, negative, or neutral).

4. Determining Sentiment

To convert these logits into readable probabilities, a 'SoftMax' function was applied, transforming the raw scores into a probability distribution for each sentiment class. Based on the probabilities, the 'argmax' function assigned the most likely sentiment category to each tweet. For instance, a tweet that scored 0.7, 0.2, and 0.1 in positive, neutral, and negative categories respectively would be labeled as positive. This process was repeated for the entire dataset.

Batch Processing and Memory Optimization

The dataset consistend of 15,000 tweets. Efficient batch processing techniques were implemented to optimize memory usage. The tweets were processed in

manageable batches, ensuring that the system's memory was not overwhelmed. A specific function, ‘classify_tweets_in_batches’, was used to divide the tweets into batches of 10 which balances memory efficiency with processing speed.

To further reduce memory usage, Python's yield functionality was employed. Unlike a typical return statement that stores all results in memory, yield allows one result to be generated at a time, minimizing RAM overhead. This approach enabled efficient classification of the tweets even with limited computational resources.

Application and Output

Upon completing the sentiment classification, sentiment labels were appended to the original dataset. Each tweet was associated with one of three categories: positive, negative, or neutral. The resulting sentiment labeled dataset was stored in a CSV format for further analysis.

This labeled dataset facilitated several types of analyses, including:

- Tracking sentiment trends over time for each company.
- Analyzing the distribution of positive, negative, and neutral tweets across different companies.
- Conducting correlation analysis to explore the relationship between sentiment differences and stock price movements.

The example table below shows a sample of sentiment-labeled tweets related to Amazon, where each row contains a tweet and its corresponding sentiment classification.

	index	post_date	text	comment_num	retweet_num	like_num	ticker_symbol	date	apple	amazon	tesla	microsoft	google	None	sentiment
0	25	2015-01-01 00:49:05	Jeff Bezos lost $7.4 billion in Amazon's worst...	21	139	57	AMZN	2015-01-01	0	1	0	0	0	0	negative
1	101	2015-01-01 03:50:06	Jeff Bezos lost $7.4 billion in Amazon's worst...	17	113	57	AMZN	2015-01-01	0	1	0	0	0	0	negative
2	229	2015-01-01 13:50:21	Analyst Report on Top Stock Market Options for...	0	430	3	AAPL	2015-01-01	1	1	0	0	0	0	neutral
3	15630	2015-01-14 12:30:00	Strong revenue growth and the prospect of impr...	2	35	16	AMZN	2015-01-14	0	1	0	0	0	0	positive
4	24525	2015-01-21 14:40:14	$AMZN Here's Why We Think Amazon's Stock Trade...	0	41	3	AMZN	2015-01-21	0	1	0	0	0	0	neutral

Figure 5: Amazon Data Frame with Sentiment Labelled Tweets

By leveraging RoBERTa's advanced sentiment analysis capabilities, this study was able to classify a large dataset of tweets into positive, negative, and neutral categories.

Stock Price Data Processing

With the tweets now sentiment-labeled, the next step involved deriving sentiment-based drivers to explore their potential influence on stock returns and transaction volumes for the six companies: Apple, Amazon, Google, Microsoft, Tesla, and Netflix. These sentiment drivers serve as inputs for assessing whether public sentiment correlates with or predicts stock price movements and trading volumes. The study developed three distinct sentiment-based drivers for further analysis:

1. Total Number of Tweets per Day (Independent of Sentiment)
2. Sentiment Differential (Positive vs. Negative Tweets)
3. Sentiment Differential Weighted by Retweets

Each of these drivers captures a different aspect of public sentiment, providing different perspectives on the relationship between social media activity and stock market behavior.

1. Total Number of Tweets per Day

The first driver measures the total number of tweets mentioning a company each day regardless of sentiment. This measure serves as a representative of overall public attention, based on the hypothesis that spikes in tweet volume may correspond to significant market developments and therefore influence stock prices and trading volumes (Bollen, Johan et al., 2011).

Calculation Process:

- The sentiment-labeled DataFrame was grouped by date, and the total number of tweets mentioning each company was summed for each trading day (2015–2019).
- A left join was performed, merging the stock price data (adjusted close prices and volumes) with the tweet data by date. This combined dataset allows for comparing stock performance with daily social media activity.

Equation 2: Total Number of Tweets per Day

$$DailyTweets = \sum_{i \in D} (Tweets_i)$$

Tracking tweet volume aims to determine whether spikes in public attention, regardless of sentiment, align with changes in stock prices. The comprehensive dataset, based on over 3.1 million tweets, ensures complete coverage with no missing sentiment data.

2. Sentiment Differential (Positive vs. Negative Tweets)

The second driver evaluates the net sentiment by calculating the difference between positive and negative tweets. This metric provides insight into whether public sentiment on a given day tends to be positive or negative for each company. A positive differential is hypothesized to correlate with rising stock prices, while a negative differential may signal potential declines (Ruiz et al., 2012).

Calculation Process:

- The sentiment-labeled DataFrame was filtered to include only positive and negative tweets, excluding neutral ones to focus on directional signals.
- The sentiment differential was calculated by subtracting the number of negative tweets from positive ones for each company on each day.
- To refine this metric, only tweets exceeding predefined engagement thresholds (e.g., a minimum number of retweets or likes) were included. For instance, Tesla tweets were filtered with a threshold of 200+ retweets, while Netflix used a lower threshold of 5+ retweets.

Equation 3: Sentiment Differential (Positive vs. Negative Tweets)

$$S_{net,d} = \sum_{i \in D} PositiveTweets_i - \sum_{i \in D} NegativeTweets_i$$

This approach ensures the sentiment differential reflects the most impactful tweets and offers a more accurate indicator of sentiment that could influence stock prices (Zhang et al., 2021).

3. Sentiment Differential Weighted by Number of Retweets

The third driver extends the sentiment differential by incorporating tweet engagement, weighting sentiment by the number of retweets. The underlying assumption is that tweets with higher engagement (e.g. retweets) have a more significant impact on public sentiment and therefore stock market behavior. Tweets retweeted hundreds or thousands of times likely reflect a broader sentiment, which may affect market activity more than tweets with minimal engagement (Sprenger et al., 2014).

Calculation Process:

- Similar to the previous step, the sentiment differential was calculated, but sentiment was weighted by the number of retweets for each tweet.
- The sentiment of each tweet (positive or negative) was multiplied by its retweet count. For example, a positive tweet with 100 retweets would have more weight than a positive tweet with only 1 retweet.

Equation 4: Sentiment Differential Weighted by Number of Retweets

$$SW_{net,d} \sum_{i \in D_+} \left(PositiveTweets_i \times Retweets_{i,pos}\right) - \sum_{i \in D_-} \left(NegativeTweets_i \times Retweets_{i,neg}\right)$$

This weighted approach emphasizes high-engagement tweets, reducing noise and better capturing sentiment that reflects broader public opinion (Tetlock, 2007).

Challenges and Missing Data

During the creation of sentiment-based drivers, challenges related to missing sentiment data on certain trading days emerged. Engagement filters were applied to exclude neutral tweets and focus on high-engagement tweets (e.g., tweets with higher numbers of retweets and likes), which was challanging. For companies with lower tweet volumes, such as Netflix, applying these filters resulted in missing sentiment data for certain trading days, as very few or no tweets met the required engagement criteria.

This reflects a trade-off between ensuring high-quality, relevant data and maintaining a complete dataset for analysis. While excluding low-engagement tweets enhances data quality by focusing on influential opinions, it also introduces gaps in the sentiment data. To address this, careful monitoring of data completeness was maintained throughout the analysis. Statistical modeling adjustments were made to minimize the impact of missing data on the final results (Little, Roderick, Rubin, 2019).

Spline Interpolation for Missing Data

To address the issue of missing sentiment data, particularly in the Daily Sentiment Differential and Weighted Sentiment Differential, spline interpolation was employed. This method provides a smooth technique for estimating missing values without introducing distortions, which is critical in maintaining the continuity required for time series analysis (Hastie et al., 2009).

Spline interpolation is a mathematical technique used to estimate missing values by fitting piecewise polynomials between known data points. Unlike higher-degree polynomial interpolation, which fits a single polynomial across the entire dataset; spline interpolation uses lower-degree polynomials, which allows for smoother transitions between points (Airbyte, 2024). In this study, a first-degree (linear) spline was used for interpolation to ensure that the general trend of the data remained complete and to avoid introducing artificial fluctuations or overfitting (Hastie et al., 2009).

The application of linear spline interpolation ensured that missing sentiment data points were filled in a manner that maintained the integrity of the time series.

Stationarity Tests and Transformation

In time series analysis, ensuring stationarity is crucial for accurate econometric modeling, particularly when performing Granger Causality tests. Non-stationary data, characterized by time-dependent statistical properties such as the mean and variance, can lead to misleading or false results in such models (Breitung and Hamilton, 1995).

There are two primary forms of stationarity:

1. Strong (Strict) Stationarity: This condition occurs when all statistical properties of the time series remain unchanged under any time shift.
2. Weak (Second-Order) Stationarity: In this form, the mean and covariance of the series remain constant over time. Weak stationarity is the more commonly required form in time series analysis, particularly for practical financial applications (Palachy, 2019).

To ensure reliable econometric analysis, both the sentiment data and stock price data must exhibit stationarity. If the data is non-stationary, relationships identified between variables may simply reflect time-dependent trends rather than true correlations (Palachy, 2019).

Augmented Dickey-Fuller (ADF) Test

To test for stationarity, the Augmented Dickey-Fuller (ADF) test was employed. The ADF test determines whether a time series has a unit root, which would indicate non-stationarity. The test follows these hypotheses:

- Null Hypothesis (H_0): The time series has a unit root, indicating non-stationarity.

- Alternative Hypothesis (H_1): The time series does not have a unit root, indicating stationarity.

If the p-value of the ADF test is below a conventional threshold (e.g., 0.05), the null hypothesis is rejected, confirming that the series is stationary (Dickey and Fuller, 1979).

Stationarity in the Data

After applying the ADF test, the sentiment-based variables, including the Daily Sentiment Differential and Weighted Sentiment Differential, along with the transaction volume data, were found to be stationary. This suggests that these variables can be directly incorporated into the econometric models without further transformation. However, the stock price data for all six companies exhibited non-stationarity, a common feature of financial time series due to trends and long-term drifts (Campbell et al., 1997). Additionally, the total daily tweet count for Apple was identified as non-stationary.

Transformation to Achieve Stationarity

To address the non-stationarity observed in the stock price data, the series was transformed using log returns. Log returns, calculated as the natural logarithmic difference between consecutive stock prices, are commonly used in financial time series analysis to stabilize variance and remove long-term trends. The formula for log returns is:

Equation 5: Log Return Transformation to Achieve Stationary

$$Log\ Return = \ln\left(\frac{Price_t}{Price_{t-1}}\right)$$

This transformation ensures that the statistical properties of the stock price data, such as its mean and variance, remain constant over time. Additionally, the daily tweet count for Apple was transformed using first differencing to achieve stationarity.

Equation 6: First Differencing Transformation for Daily Tweet Count to Achieve Stationary

$$\Delta\ tweet_t = \log\left(\frac{\sum_{i\in D}(Tweets_i)}{\sum_{i\in D}(Tweets_{i-1})}\right)$$

These transformations prepared the dataset for further analysis, ensuring that the data was stationary and suitable for methods like the Granger Causality Test. This minimized the risk of generating false results due to non-stationary data.

Granger Causality

With the data cleaned, interpolated, and transformed to ensure stationarity, the next step was to test the predictive power of sentiment data on stock returns and transaction volumes using Granger causality.

Overview of Granger Causality

Developed by Clive Granger in 1969, Granger causality is a statistical method used to determine whether one time series can be used to predict another (Granger, 1969). This approach is widely utilized in econometrics and time series analysis to assess whether a predictive relationship exists between two variables. The fundamental concept behind Granger causality is that if one time series (denoted as X) contains useful information for forecasting the future values of another time series (denoted as Y), after accounting for Y's own past values, then X is said to "Granger-cause" Y (Granger, 1969).

It is important to note that Granger causality does not imply a direct causal relationship in the traditional sense (Granger, 1969). Rather, it indicates that there is a statistically significant predictive relationship between the two time series.

The steps for conducting the Granger causality test in this study were as follows:

1. Fit an Ordinary Least Squares (OLS) model using the past values of the stock returns (or transaction volumes) to predict their future values.
2. Add the past values of the sentiment indicators (Daily Sentiment Differential or Weighted Sentiment Differential) to the model and evaluate whether the prediction improves.
3. If the inclusion of the past sentiment indicators significantly improves the prediction of future stock returns (or transaction volumes), then sentiment indicators are said to "Granger-cause" stock returns (or transaction volumes).

Additionally, to test the reverse relationship:

1. Fit an Ordinary Least Squares (OLS) model using the past values of the sentiment indicators to predict their future values.
2. Add the past values of the stock returns (or transaction volumes) to the model and evaluate whether the prediction improves.

3. If the inclusion of the past stock returns (or transaction volumes) significantly improves the prediction of future sentiment indicators, then stock returns (or transaction volumes) are said to "Granger-cause" sentiment indicators.

Bidirectional Analysis

To explore the relationship between sentiment and stock market behavior, the Granger causality test was conducted in both directions. This bidirectional analysis allowed for an examination of whether sentiment influences stock returns and whether stock returns influence public sentiment. This approach provides a more complete understanding of the dynamic interactions between social media sentiment and market behavior (Bollen, Johan et al., 2011).

Cross-Stock Comparison

The Granger causality tests were conducted individually for each of the six stocks to determine if predictive relationships observed in one stock were consistent across the others. This cross-stock comparison aimed to identify patterns that could be generalized beyond individual stocks to the broader category of popular technology companies. Such insights are valuable for understanding whether the findings for companies like Tesla and Apple could apply to a wider range of firms in similar sectors (Zhang et al., 2021).

Significance Testing

The ‘grangercausalitytests’ function from the statsmodels library in Python was used to perform hypothesis testing and determine whether there was sufficient evidence to reject the null hypothesis of no Granger causality. A low p-value (typically less than 0.05) suggests that one time series Granger-causes the other. Each test was evaluated to ensure that the predictive relationships detected were statistically significant.

Through the application of spline interpolation, stationarity testing, and Granger causality, a strong dataset suitable for time series analysis was created. Additionally, the exploration of whether stock market behavior influences public sentiment provided a two-way analysis, offering a framework for understanding the dynamic relationship between social media sentiment and financial markets.

INTERPRETATION OF RESULTS

Granger Causality: Total Daily Sentiment vs. Stock Returns and Transaction Volume

The results from the Granger causality tests between the total daily number of tweets and stock returns and transaction volume are presented in 'Figure 6: Granger Causality: Total Daily Sentiment vs. Stock Returns and Transaction Volume'. The analysis desired to determine whether sentiment (as indicated by the total number of tweets) can predict stock returns or transaction volumes for the six companies: Apple, Amazon, Google, Microsoft, Tesla, and Netflix.

Stock Returns Prediction

The results indicate that for Microsoft, total daily sentiment Granger-causes stock returns at a significance level of 5%, specifically for the first two lags. This suggests that spikes in social media activity for Microsoft may precede changes in its stock returns, though the relationship is not consistent across all lags.

In contrast, for the other five companies, there is no consistent evidence that total sentiment Granger-causes stock returns at the 5% significance level. This suggests that, for these companies, overall tweet volume does not predict stock returns. Apple, Amazon, Google, Tesla, and Netflix show p-values above the 5% threshold across most lags, indicating a lack of predictive power for sentiment in driving their stock returns.

More interestingly, the reverse Granger causality was observed, where stock returns predict the total number of tweets. In four out of six companies, there is significant evidence that the first three or four lags of stock returns predict total sentiment (measured by tweet volume) at the 5% level. This reverse causality suggests that fluctuations in stock returns are followed by changes in tweet volume. In other words, rather than social media sentiment driving stock returns, it appears that changes in stock performance spur public interest and increase tweet activity.

Transaction Volume Prediction

The effect of transaction volume on total sentiment stands out. Across all six companies, transaction volume Granger-causes total sentiment, with statistical significance at 5% for most lags. This indicates that trading activity in these stocks is a

significant driver of social media discussions. Spikes in transaction volume may lead to increased attention on platforms like Twitter, reflecting the market's overall interest and engagement with these companies.

However, when examining the reverse, whether total sentiment predicts transaction volume, there is no statistically significant Granger causality for five out of six companies (excluding Tesla). This suggests that while transaction volume strongly influences tweet volume, the reverse is not true; tweet volume alone is not a reliable predictor of trading activity in most cases. These findings point to an important dynamic: transaction volume seems to play a more critical role in driving public attention (as measured by tweet volume) than sentiment cases does in driving transaction volume or stock returns. The direction of the predictability relationship is thus: transaction volume to total sentiment.

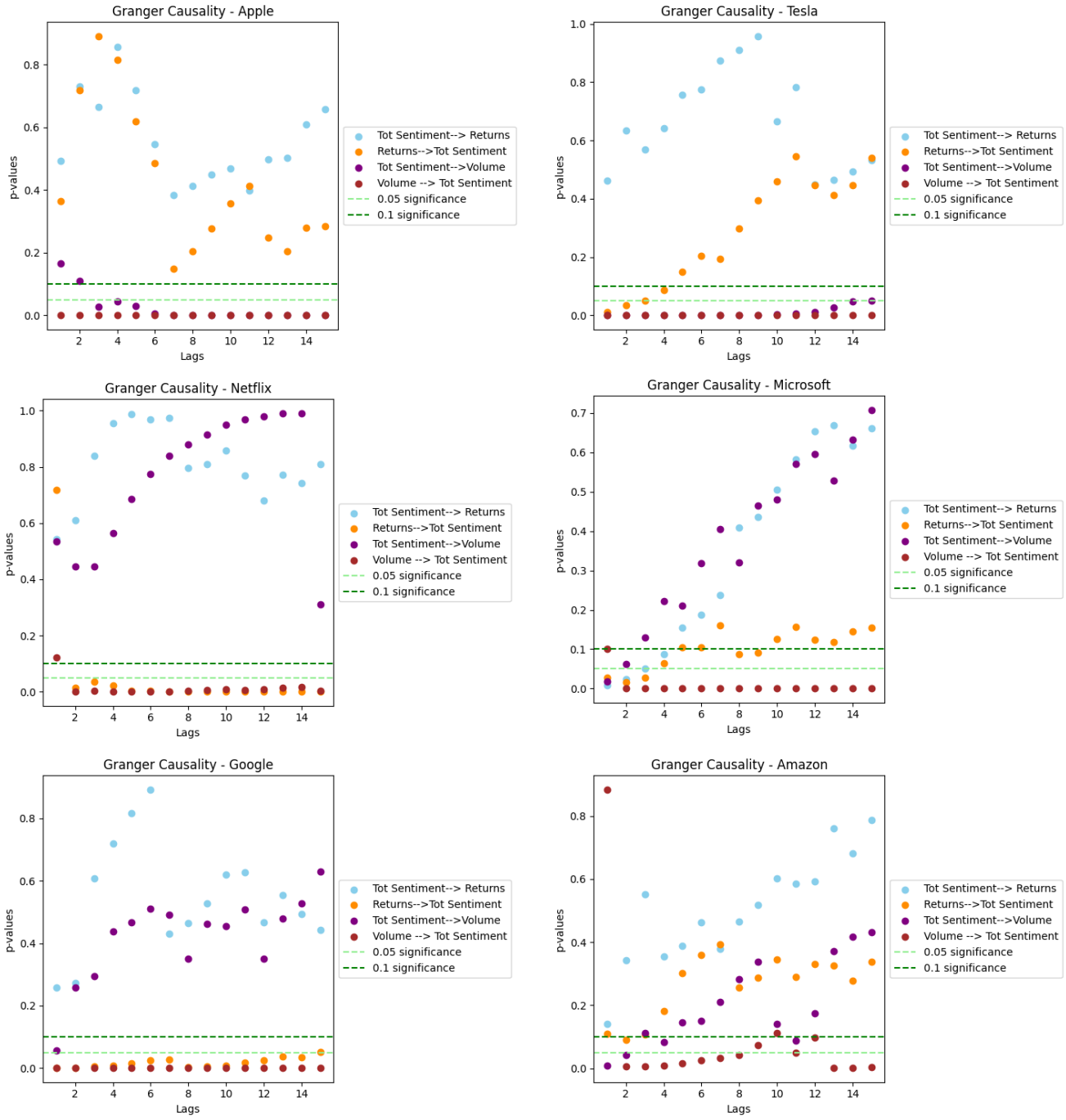

Figure 6: Granger Causality: Total Daily Sentiment vs. Stock Returns and Transaction Volume

Granger Causality: Sentiment Differential (Unweighted & Weighted) vs. Stock Returns

Further analysis focused on the relationship between the sentiment differential (both weighted by retweets and unweighted) and stock returns, as shown in 'Figure 7: Granger Causality: Total Daily Sentiment unweighted & weighted vs. Stock Returns'. This analysis aimed to identify whether the net sentiment (positive minus negative tweets) can predict stock returns and, conversely, whether stock returns predict sentiment shifts.

Stock Returns Prediction

The results show that for all six companies, the sentiment differential Granger-causes stock returns at 5% significance for at least one lag. Furthermore, Tesla, Google, Netflix and Amazon demonstrate that the sentiment differential Granger-causes stock returns across multiple lags. This consistent pattern highlights that shifts in net sentiment, where there are more positive tweets than negative ones, can predict future stock returns for the majority of these companies.

When examining the retweets weighted sentiment differential, the findings are slightly more subtle. The weighted sentiment Granger causes stock returns at 10% significance for four out of six companies, with Microsoft and Netflix as exceptions. The reduced number of datapoints for Netflix and Microsoft (approximately 1000 compared to 4000 for the other companies) may have influenced these results. The lower volume of high engagement tweets for these stocks suggests that the influence of sentiment on returns may be less strong for companies with less social media engagement.

Reverse Granger Causality

The reverse Granger causality results indicate that stock returns Granger-cause sentiment differentials in five out of six companies at 5% significance for unweighted sentiment with Microsoft as exception. The reverse Granger causality results indicate that stock returns Granger-cause weighted sentiment differentials at 5% significance for companies Apple, Amazon, Google. This finding suggests that stock returns influence public sentiment, where positive stock performance may lead to a higher volume of positive tweets, thereby shifting the sentiment differential.

However, this relationship is not as strong for the weighted sentiment differential, where only three out of six companies demonstrate significant reverse causality. This result indicates that while stock performance can influence general sentiment (unweighted), it may not have as strong an impact on high-engagement tweets.

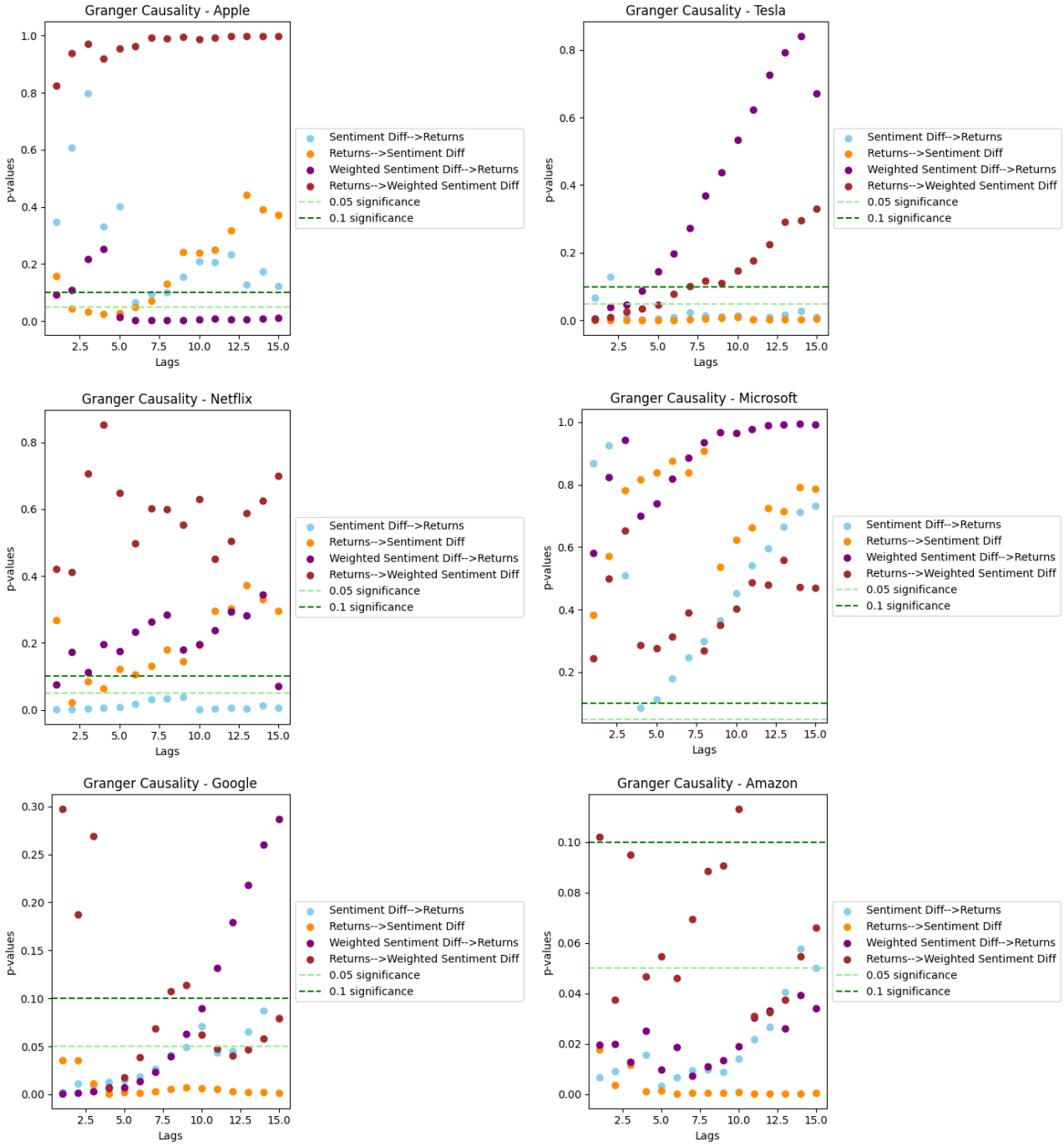

Figure 7: Granger Causality: Total Daily Sentiment unweighted & weighted vs. Stock Returns

FINDINGS AND CONSIDERATIONS

Key Finding 1: Transaction Volume's Influence on Social Media Sentiment

The most notable finding concerns to the Granger causality between the total daily number of tweets and transaction volumes. The dataset, which averaged around 300,000 tweets per stock, provided high reliability in the results. The Granger causality tests showed that transaction volume Granger-caused the total number of tweets, but the reverse relationship was not found to be statistically significant. This indicates that public engagement on social media platforms such as Twitter is driven more by market trading activity rather than the other way around.

This finding suggests that market participants respond more to trading volumes than to price changes. Notably, this insight applies more to broad social media platforms like Twitter, where many users may not be directly involved in trading activities, and thus public attention is reflective of trading activity rather than price fluctuations.

Key Finding 2: Net Sentiment's Predictive Power Over Stock Returns

The second significant finding is the predictive power of net sentiment (i.e., the difference between positive and negative tweets) over stock returns. The analysis showed that net sentiment, particularly when weighted by retweets, Granger-causes stock returns across the majority of the stocks examined. This implies that shifts in public opinion on social media can serve as a meaningful indicator of stock performance.

The reverse relationship, where stock returns Granger-caused net sentiment, was observed in fewer cases, further confirming the stronger predictive capacity of sentiment on stock returns. This asymmetry aligns with existing research, reinforcing the role of public sentiment as a leading indicator for stock market performance.

Considerations and Limitations

While this study provides valuable insights, several limitations must be considered. The dataset was confined to tweets from 2015 to 2020 and focused on only six companies, all from the technology sector. Thus, the findings may not be generalizable to other industries. Additionally, filtering tweets based on engagement metrics limited the dataset, particularly for stocks like Netflix and Microsoft, which had lower overall tweet

volumes. The results for these two companies should be viewed cautiously, as the reduced sample size may affect the reliability of the analysis

Furthermore, the RoBERTa model, while effective in sentiment classification, has its limitations in interpreting informal social media language, such as sarcasm or ambiguous statements. Misclassification of sentiment could have affected the results, and more advanced models may yield even more accurate classifications.

Lastly, while measures were taken to ensure stationarity, the assumption of stationarity in financial data remains challenging. The complexity of financial markets may introduce trends that are difficult to capture fully, even with transformations such as log returns.

Managerial Implications

The findings have practical implications for financial professionals, particularly those focused on developing data-driven trading strategies. By incorporating sentiment analysis into quantitative models, traders could gain a competitive advantage by using public sentiment as an additional layer of information to predict stock movements. Real-time tools that track shifts in public sentiment could be highly useful in detecting changes that might signal future stock price movements.

Furthermore, the insight that transaction volume drives social media activity adds an extra dimension to trading strategies. Financial institutions, particularly those involved in algorithmic trading or portfolio management, may benefit from integrating sentiment data into their existing prediction models.

Comparison to Existing Literature

This study aligns with previous research that explores the role of social media sentiment in financial markets but provides a broader perspective by examining public sentiment rather than influencer-driven sentiment shifts. Unlike previous studies focused on the impact of high-profile individuals like Elon Musk, this research analyzed aggregate sentiment across multiple stocks, which offers a more generalized understanding of how sentiment influences market performance.

Future Research Directions

Future studies could expand upon these findings by:

1. Including a larger dataset with more companies across various sectors and geographical markets to better understand how sentiment impacts different industries.
2. Integrating other social media platforms such as Reddit or Facebook to explore sentiment in specifically trading communities like WallStreetBets.
3. Exploring advanced machine learning models to further refine sentiment classification accuracy.
4. Backtesting sentiment-based trading strategies under varying market conditions to assess their effectiveness across different market environments.

This research contributes to understanding the dynamic relationship between social media sentiment and financial markets, laying the groundwork for further exploration in this evolving field.

BIBLIOGRAPHY

- Akin, Isik, and Meryem Akin. “Behavioral Finance Impacts on US Stock Market Volatility: An Analysis of Market Anomalies.” Behavioural Public Policy, 13 Mar. 2024, pp. 1–25, https://doi.org/10.1017/bpp.2024.13.

- Airbyte, “Data Interpolation: What It Is & How to Do It?” *Airbyte.com*, May 2024, airbyte.com/data-engineering-resources/data-interpolation.

- Alvarez, Emiliano, et al. “The Dynamical Relation between Price Changes and Trading Volume.” *Quality & Quantity*, 10 Jan. 2023, https://doi.org/10.1007/s11135-022-01605-4.

- Bailey, Liam D. “Dr. Liam D. Bailey.” *Liamdbailey.com*, 30 May 2022, liamdbailey.com/posts/2022-05-30-missing-data-different-interpolation-methods/.

- Bhadamkar, Amey, and Sonali Bhattacharya. “Tesla Inc. Stock Prediction Using Sentiment Analysis.” *AABFJ*, vol. 16, 2022, p. 52.

- Bodie, Z., Kane, A. and Marcus, A. (2014) Investments. Global Edition, 10th Edition, McGraw Hill Higher Education, New York.

- Bollen, Johan, et al. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science*, vol. 2, no. 1, Mar. 2011, pp. 1–8.

- Breitung, Jorg, and James D. Hamilton. “Time Series Analysis.” *Contemporary Sociology*, vol. 24, no. 2, Mar. 1995, p. 271, https://doi.org/10.2307/2076916.

- Campbell, John, et al. *The Econometrics of Financial Markets*. 1997.

- Delong, J, and Konstantin Magin. *NBER WORKING PAPER SERIES a SHORT NOTE on the SIZE of the DOT-COM BUBBLE*. 2006.

- Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *ArXiv.org*, 11 Oct. 2018, arxiv.org/abs/1810.04805.

- Dickey, David A., and Wayne A. Fuller. “Distribution of the Estimators for Autoregressive Time Series with a Unit Root.” *Journal of the American Statistical Association*, vol. 74, no. 366, June 1979, pp. 427–431, www.jstor.org/stable/2286348, https://doi.org/10.2307/2286348.

- Dogan, Mustafa, et al. “Speculator and Influencer Evaluation in Stock Market by Using Social Media.” *2020 IEEE International Conference on Big Data (Big Data)*, 10 Dec. 2020, pp. 4559–4566, https://doi.org/10.1109/bigdata50022.2020.9378170.

- Gossett, Stephen. “What Is Alternative Data and Why Is It Changing Finance?” *Built In*, Dec. 2023, builtin.com/articles/alternative-data.

- Granger, C. W. J. “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods.” *Econometrica*, vol. 37, no. 3, Aug. 1969, pp. 424–438, https://doi.org/10.2307/1912791.

- Gratton, Peter. “What Is the Stock Market and How Does It Work?” *Investopedia*, 12 Mar. 2022, www.investopedia.com/terms/s/stockmarket.asp.

- Grigory Sapunov. “[Papers] RoBERTa: A Robustly Optimized BERT Pretraining Approach.” Medium, Medium, 29 July 2019, moocaholic.medium.com/papers-roberta-a-robustly-optimized-bert-pretraining-approach-7449bc5423e7. Accessed 17 Sept. 2024.

- Guo, Kun, et al. “Can Investor Sentiment Be Used to Predict the Stock Price? Dynamic Analysis Based on China Stock Market.” *Physica A: Statistical Mechanics and Its Applications*, vol. 469, Mar. 2017, pp. 390–396, https://doi.org/10.1016/j.physa.2016.11.114.

- Hayes, Adam. “Dividend Definition.” *Investopedia*, 17 Dec. 2023, www.investopedia.com/terms/d/dividend.asp.

- Hastie, Trevor, et al. *The Elements of Statistical Learning*. *Springer Series in Statistics*, New York, NY, Springer New York, 2009, link.springer.com/book/10.1007/978-0-387-84858-7.

- Howard, Jeremy, and Sebastian Ruder. “Universal Language Model Fine-Tuning for Text Classification.” *ArXiv.org*, 2018, arxiv.org/abs/1801.06146.

- Jha, Saurabh. “BERT Pre-Training: - Saurabh Jha - Medium.” Medium, Medium, 5 Mar. 2024, medium.com/@saurabhjha443/bert-pre-training-simplified-7a8872b1cce9.

- JP Morgan, *Big Data and AI Strategies 2019 Alternative Data Handbook Global Quantitative and Derivatives Strategy*. 11 Oct. 2019.

- Kalampokis, Evangelos, et al. “Understanding the Predictive Power of Social Media.” *Internet Research*, vol. 23, no. 5, 14 Oct. 2013, pp. 544–559, https://doi.org/10.1108/intr-06-2012-0114.

- Little, Roderick, and Donald Rubin. *Statistical Analysis with Missing Data, Third Edition*. Wiley, 2019.

- Liu, Yinhan, et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *ArXiv.org*, 26 July 2019, arxiv.org/abs/1907.11692.

- Metta, Sanjeev, et al. “Power of 280: Measuring the Impact of Elon Musk’s Tweets on the Stock Market.” *Ushus Journal of Business Management*, vol. 21, no. 1, 3 July 2022, pp. 17–43, https://doi.org/10.12725/ujbm.58.2.

- Mittal, Anshul, and Arpit Goel. *Stock Prediction Using Twitter Sentiment Analysis*. 2017

- Palachy, Shay. “Stationarity in Time Series Analysis.” *Towards Data Science*, Towards Data Science, 8 Apr. 2019, towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322.

- Ruiz, Eduardo J., et al. “Correlating Financial Time Series with Micro-Blogging Activity.” *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, 8 Feb. 2012, https://doi.org/10.1145/2124295.2124358.

- Ranco, Gabriele, et al. “The Effects of Twitter Sentiment on Stock Price Returns.” *PLOS ONE*, vol. 10, no. 9, 21 Sept. 2015, p. e0138441, https://doi.org/10.1371/journal.pone.0138441.

- Sprenger, Timm & Tumasjan, Andranik & Sandner, Philipp & Welpe, Isabell. (2014). Tweets and Trades: the Information Content of Stock Microblogs. European Financial Management. 20. 926-957. 10.1111/j.1468-036X.2013.12007.x.

- Tetlock, Paul. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *The Journal of Finance*, vol. 62, no. 3, 8 May 2007, pp. 1139–1168.

- Weishaupt, Martin. *Predicting Tesla Stock Return Using Twitter Data an Intraday View on the Relation between Twitter Dimensions and the Tesla Stock Return By*. 2020.

- Zhang, Weiguo, et al. “Predicting Stock Market Volatility Based on Textual Sentiment: A Nonlinear Analysis.” *Journal of Forecasting*, vol. 40, no. 8, 4 May 2021, https://doi.org/10.1002/for.2777.

APPENDIX

DATASET:

Tweets - Top Companies from 2015 to 2020 (kaggle.com)

LINK TO GOOGLE COLAB NOTEBOOK:

Final_Thesis_Code_ROBERTA.ipynb

Final_Thesis_Code.ipynb

LINK TO JUPYTER NOTEBOOK :

http://localhost:8888/notebooks/Thesis%20-%20Worksheet.ipynb

Granger Causality Test Results for Sentiment and Stock Returns with Weighted and Unweighted Sentiment Differentials (15 Lags):

GOOGLE:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.002492	0.035393	0.000528	0.297547
1	2	0.011304	0.035658	0.001790	0.187146
2	3	0.009462	0.011225	0.003608	0.269365
3	4	0.013111	0.000767	0.007413	0.006115
4	5	0.014366	0.002420	0.007497	0.017827
5	6	0.018457	0.001717	0.013374	0.039153
6	7	0.026903	0.002893	0.023623	0.068476
7	8	0.041361	0.005842	0.039849	0.107460
8	9	0.049271	0.007508	0.063418	0.114061
9	10	0.071182	0.006502	0.089422	0.062496
10	11	0.043292	0.006040	0.131630	0.047867
11	12	0.045358	0.003011	0.179727	0.040577
12	13	0.065082	0.002119	0.217959	0.047070
13	14	0.086883	0.002742	0.259862	0.057935
14	15	0.079955	0.001389	0.286434	0.079379

AMAZON:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.006652	0.017774	0.019659	0.102054
1	2	0.009039	0.003496	0.019785	0.037418
2	3	0.012527	0.011583	0.012857	0.094892
3	4	0.015649	0.001007	0.025038	0.046680
4	5	0.003363	0.001417	0.009865	0.054626
5	6	0.006694	0.000219	0.018707	0.045983
6	7	0.009510	0.000377	0.007165	0.069480
7	8	0.009732	0.000423	0.011006	0.088455
8	9	0.008876	0.000434	0.013581	0.090794
9	10	0.013974	0.000928	0.019024	0.113231
10	11	0.021784	0.000106	0.030405	0.030982
11	12	0.026628	0.000126	0.033096	0.032409
12	13	0.040545	0.000151	0.026159	0.037491
13	14	0.057713	0.000307	0.039301	0.054653
14	15	0.050095	0.000382	0.034095	0.066116

MICROSOFT:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.867790	0.382137	0.581060	0.243585
1	2	0.926402	0.571868	0.823620	0.498800
2	3	0.508614	0.781147	0.941905	0.653195
3	4	0.085155	0.816299	0.699102	0.287056
4	5	0.113469	0.837600	0.739667	0.277602
5	6	0.179248	0.874769	0.819205	0.312586
6	7	0.246749	0.838135	0.886331	0.391337
7	8	0.299719	0.907664	0.935825	0.268791
8	9	0.366294	0.535525	0.966437	0.349757
9	10	0.451848	0.622314	0.965139	0.402289
10	11	0.542739	0.662659	0.978665	0.488228
11	12	0.596949	0.725911	0.989206	0.479161
12	13	0.665335	0.715190	0.991777	0.559301
13	14	0.713446	0.791294	0.995560	0.471669
14	15	0.731814	0.786583	0.991442	0.470011

NETFLIX:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.000547	0.267423	0.076067	0.419881
1	2	0.000974	0.021904	0.172177	0.411210
2	3	0.003161	0.085023	0.113421	0.705863
3	4	0.005305	0.064919	0.196627	0.852952
4	5	0.009385	0.122792	0.175637	0.648455
5	6	0.018534	0.105146	0.232380	0.498503
6	7	0.032219	0.131850	0.263484	0.602230
7	8	0.032938	0.179739	0.284058	0.599124
8	9	0.039031	0.143997	0.179933	0.553789
9	10	0.002058	0.193045	0.195496	0.629779
10	11	0.003893	0.295235	0.238860	0.450388
11	12	0.006202	0.303698	0.294602	0.504580
12	13	0.004026	0.371659	0.282109	0.589012
13	14	0.011974	0.330165	0.343958	0.624897
14	15	0.006011	0.295709	0.071082	0.698243

TESLA:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.067623	0.000058	0.004846	0.003078
1	2	0.128165	0.000168	0.040034	0.009441
2	3	0.009279	0.000311	0.046198	0.025514
3	4	0.003432	0.000591	0.088654	0.034704
4	5	0.006319	0.001016	0.144054	0.045710
5	6	0.009808	0.001731	0.197866	0.077733
6	7	0.023391	0.003758	0.272162	0.102004
7	8	0.014029	0.005586	0.369670	0.118185
8	9	0.011589	0.007475	0.438402	0.111300
9	10	0.015630	0.010710	0.533494	0.147873
10	11	0.004086	0.002285	0.623177	0.176229
11	12	0.010538	0.002399	0.726057	0.225967
12	13	0.016513	0.003265	0.791632	0.290435
13	14	0.027680	0.003048	0.840946	0.296234
14	15	0.009308	0.004587	0.671627	0.329411

APPLE:

	Lag	p_values_S_R	p_values_R_S	p_values_S_R_RETweight	p_values_R_S_RETweight
0	1	0.348832	0.158849	0.093978	0.823619
1	2	0.609079	0.043875	0.108941	0.937170
2	3	0.796267	0.033226	0.218753	0.971702
3	4	0.331014	0.025742	0.252143	0.918908
4	5	0.401076	0.026767	0.013521	0.954444
5	6	0.064862	0.050640	0.003116	0.964198
6	7	0.095984	0.072080	0.002444	0.991670
7	8	0.100655	0.132047	0.003759	0.991202
8	9	0.155312	0.242829	0.004135	0.994824
9	10	0.208677	0.239170	0.006780	0.986036
10	11	0.205558	0.251167	0.010101	0.993057
11	12	0.235107	0.317659	0.006907	0.996787
12	13	0.128793	0.441486	0.006494	0.998233
13	14	0.173282	0.391349	0.009987	0.998907
14	15	0.122592	0.371607	0.010209	0.998059

Granger Causality Test Results for Sentiment, Stock Returns, and Trading Volume (15 Lags):

APPLE:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.493022	0.364459	0.164571	2.097579e-05
1	2	0.729488	0.719536	0.111081	1.670153e-15
2	3	0.665854	0.891366	0.028547	2.915362e-16
3	4	0.856446	0.814822	0.045182	1.141691e-15
4	5	0.718214	0.618590	0.030355	1.536584e-14
5	6	0.546689	0.485182	0.006843	1.130904e-14
6	7	0.384485	0.149091	0.000039	4.000288e-14
7	8	0.412866	0.205123	0.000112	1.141079e-13
8	9	0.449147	0.277809	0.000045	7.102585e-14
9	10	0.468866	0.357020	0.000195	1.869725e-13
10	11	0.398734	0.413093	0.000211	3.482232e-13
11	12	0.497594	0.247118	0.000352	1.009056e-12
12	13	0.501870	0.204935	0.000416	1.843502e-12
13	14	0.610297	0.280661	0.000721	4.401466e-12
14	15	0.656804	0.284974	0.001073	1.178999e-11

TESLA:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.462075	0.011438	3.748080e-18	2.083753e-04
1	2	0.634630	0.033952	1.661884e-13	2.433745e-09
2	3	0.570051	0.050620	1.461624e-10	1.368234e-09
3	4	0.643202	0.085948	1.026126e-08	3.154640e-09
4	5	0.756497	0.148209	2.935171e-07	8.648724e-09
5	6	0.775145	0.203054	3.735495e-05	3.457929e-09
6	7	0.873783	0.192901	2.524885e-04	1.336266e-08
7	8	0.910570	0.298725	4.205396e-04	4.573572e-08
8	9	0.957954	0.393100	2.143465e-03	7.341646e-08
9	10	0.664334	0.459644	2.426445e-03	7.320933e-08
10	11	0.782393	0.544438	7.356516e-03	4.329080e-07
11	12	0.448285	0.446181	1.183166e-02	3.046402e-07
12	13	0.465049	0.411687	2.754091e-02	3.963184e-07
13	14	0.494180	0.446920	4.653821e-02	9.644202e-07
14	15	0.531711	0.541401	4.961485e-02	1.155222e-06

GOOGLE:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.257852	0.001810	0.057324	3.751260e-11
1	2	0.272816	0.001219	0.259451	5.905783e-15
2	3	0.607631	0.004493	0.294749	9.123524e-15
3	4	0.719033	0.009243	0.438261	1.901504e-15
4	5	0.816446	0.015349	0.466613	2.325428e-15
5	6	0.892834	0.025620	0.509780	1.116986e-14
6	7	0.430376	0.028377	0.491689	1.972752e-14
7	8	0.464324	0.004414	0.349723	6.530823e-14
8	9	0.527691	0.005171	0.463365	6.317627e-14
9	10	0.619686	0.009297	0.453857	9.671457e-14
10	11	0.628238	0.017204	0.508798	7.806360e-14
11	12	0.466418	0.024503	0.351200	1.463319e-13
12	13	0.553511	0.038068	0.478181	9.364650e-14
13	14	0.494981	0.035434	0.527587	2.414050e-13
14	15	0.442912	0.052164	0.629131	3.127468e-13

AMAZON:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.140946	0.110973	0.009318	0.883864
1	2	0.343387	0.091355	0.043668	0.006165
2	3	0.552190	0.108573	0.111147	0.006824
3	4	0.354905	0.181716	0.083364	0.009331
4	5	0.389332	0.302655	0.146637	0.017438
5	6	0.463160	0.360634	0.151410	0.025068
6	7	0.378811	0.392569	0.211935	0.032111
7	8	0.464712	0.255382	0.283507	0.043686
8	9	0.517687	0.287694	0.338024	0.073342
9	10	0.602326	0.345496	0.140116	0.112204
10	11	0.585784	0.289269	0.088353	0.050289
11	12	0.591637	0.331960	0.175609	0.098752
12	13	0.761421	0.327166	0.372840	0.001051
13	14	0.681112	0.278196	0.418073	0.001741
14	15	0.787832	0.338951	0.431822	0.003627

NETFLIX:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.542465	0.718599	0.534904	0.122350
1	2	0.609252	0.013641	0.446574	0.000865
2	3	0.839714	0.035462	0.444492	0.002642
3	4	0.955041	0.021336	0.564948	0.000394
4	5	0.988684	0.003825	0.686390	0.000372
5	6	0.969251	0.004676	0.773517	0.000729
6	7	0.974834	0.000176	0.839220	0.001663
7	8	0.796050	0.000075	0.878729	0.003576
8	9	0.810246	0.000178	0.915186	0.007601
9	10	0.857394	0.000403	0.950973	0.009147
10	11	0.768608	0.000556	0.969721	0.005648
11	12	0.679453	0.000973	0.980736	0.009258
12	13	0.772414	0.000791	0.990554	0.015131
13	14	0.741525	0.000881	0.990969	0.017433
14	15	0.810915	0.001493	0.310379	0.002855

MICROSOFT:

	Lag	p_values_S_R	p_values_R_S	p_values_S_V	p_values_V_S
0	1	0.008021	0.027170	0.017422	1.008533e-01
1	2	0.023228	0.016825	0.061671	1.068217e-07
2	3	0.051068	0.027599	0.129550	1.288041e-07
3	4	0.087026	0.064666	0.221450	5.899373e-07
4	5	0.155282	0.104910	0.210539	8.497673e-07
5	6	0.187333	0.104763	0.318477	2.166589e-06
6	7	0.238203	0.160505	0.405617	3.591650e-06
7	8	0.409443	0.088134	0.320381	4.142671e-06
8	9	0.435121	0.091168	0.464412	5.924815e-06
9	10	0.505567	0.125566	0.479753	2.640522e-06
10	11	0.582670	0.157006	0.570068	1.817837e-06
11	12	0.653619	0.124462	0.595742	3.389507e-06
12	13	0.669860	0.118194	0.528123	4.857044e-06
13	14	0.617566	0.145525	0.633198	7.441145e-06
14	15	0.662116	0.155232	0.708247	9.767886e-06