# Exploring the Impact of Financial News on Stock Prices: A Knowledge Graph and Causal Inference Approach

## 1 Introduction

In this report, I investigate the relationship between financial news and stock price movements using a combination of knowledge graph construction and causal inference techniques. The focus is on analyzing how Apple Inc.'s stock prices are influenced by the financial news cycle, leveraging web-scraped data, natural language processing (NLP), and Granger causality testing.

### 1.1 Research Motivation

Financial markets are sensitive to public sentiment and information flow, as reflected in stock prices. Understanding the causal relationship between news sentiment and stock price volatility can provide valuable insights for investors and market analysts. However, accurately determining causality remains a challenge due to the complexity of financial systems and the unstructured nature of news data.

### 1.2 Research Questions

1. Can financial news sentiment be quantitatively linked to stock price fluctuations?
2. Is there a causality between published news articles and the subsequent movement in Apple's stock prices?

## 2 Data Collection

The project involved collecting two primary datasets: financial news articles about Apple Inc. and its corresponding stock prices for the period of July

2024.

## 2.1 Web Scraping for News Articles

I used the NewsAPI to collect articles related to Apple Inc., focusing on the date range from July 1, 2024, to July 31, 2024. The following Python code was employed to extract and store articles, ensuring the dataset was comprehensive for analysis.

```python
import requests
import pandas as pd
api_key = 'your_api_key'
query = 'Apple Inc.'
time_ranges = [
    ('2024-07-14', '2024-07-14'),
    ('2024-07-15', '2024-07-15'),
    ...
]
page_size = 100
all_articles = []
for start_date, end_date in time_ranges:
    page = 1
    while True:
        news_url = (
        f'https://newsapi.org/v2/everything?q={query}&from={start_date}&to={end_date}&'
        f'sortBy=publishedAt&language=en&pageSize={page_size}&page={page}&apiKey={api_key}'
        )
        response = requests.get(news_url)
        data = response.json()
        if 'articles' in data and data['articles']:
            articles = data['articles']
            all_articles.extend(articles)
            if len(articles) < page_size:
                break
            page += 1
        else:
            break
news_df = pd.DataFrame(all_articles)
news_df.to_csv('apple_news_data.csv', index=False)
```

Listing 1: Web Scraping for News Articles

## 2.2 Stock Price Data Collection

I used the Yahoo Finance API to retrieve historical stock price data for Apple (AAPL), focusing on daily close prices. The data covers the same period as the news articles for coherence in subsequent analyses.

```python
import yfinance as yf
def fetch_stock_data(symbol, period=’1y’):
    stock = yf.Ticker(symbol)
    data = stock.history(period=period)
    return data
stock_data = fetch_stock_data("AAPL")
if not stock_data.empty:
    stock_data.to_csv(’apple_stock_data.csv’)
else:
    print("No stock data fetched.")
```

Listing 2: Stock Price Data Collection

# 3 Data Preprocessing

Both news and stock price data required preprocessing for consistent analysis. The news dataset was cleaned, focusing on article content and publication dates, while the stock prices were aligned by date.

## 3.1 Date Normalization

To ensure both datasets were comparable, I normalized the date format and ensured uniformity across time zones. Articles published on the same day as stock price fluctuations were considered part of the same temporal window.

```python
news_data[’publishedAt’] = pd.to_datetime(news_data[’publishedAt’], errors=’coerce’, utc=True)
stock_data[’Date’] = pd.to_datetime(stock_data[’Date’], errors=’coerce’, utc=True)
```

Listing 3: Date Normalization

## 3.2 Sentiment Analysis

I performed sentiment analysis on the news articles using the TextBlob library to quantify the emotional tone of each article. This provided a sentiment score for each article, ranging from -1 (negative) to 1 (positive).

```python
from textblob import TextBlob
news_data['sentiment'] = news_data['content'].apply(lambda x: TextBlob(x).sentiment.polarity if pd.notnull(x) else 0)
```

Listing 4: Sentiment Analysis

## 3.3 Entity Extraction with spaCy

Named Entity Recognition (NER) was performed using spaCy to extract key entities (such as organizations, persons, and events) from the news content. This allowed for building relationships between entities within a knowledge graph.

```python
import spacy
nlp = spacy.load('en_core_web_sm')
entities = set()
for doc in nlp.pipe(news_data['content']):
    for ent in doc.ents:
        entities.add((ent.text, ent.label_))
```

Listing 5: Entity Extraction with spaCy

# 4 Knowledge Graph Construction

I utilized the Neo4j graph database to store entities and relationships extracted from the news articles. The knowledge graph enabled structured representation and querying of relationships between entities.

## 4.1 Graph Creation

The nodes in the knowledge graph represent entities such as companies (e.g., Apple Inc.), people, and products, while the edges represent relationships like acquisition, product releases, or partnerships.

```python
from py2neo import Graph, Node, Relationship
graph = Graph("bolt://localhost:7687", auth=("neo4j", "password"))
for entity in entities:
    node = Node(entity[1], name=entity[0])
    graph.merge(node, entity[1], "name")
```

Listing 6: Graph Creation with Neo4j

## 4.2 Relation Extraction

I extracted subject-verb-object relationships from the text using dependency parsing in spaCy. These relations were then added to the Neo4j knowledge graph to link entities together.

```python
def extract_relations(doc):
    relations = []
    for token in doc:
        if token.dep_ == "nsubj" and token.head.pos_ ==
            "VERB":
            subject = token.text
            verb = token.head.lemma_.upper()
            obj = None
            for child in token.head.children:
                if child.dep_ in ("dobj", "attr", "prep"
                    ):
                    obj = child.text
                    break
            if obj:
                relations.append((subject, verb, obj))
    return relations
```

Listing 7: Relation Extraction

# 5 Causal Inference Analysis

To investigate the potential causal relationship between news sentiment and stock prices, I performed a Granger causality test. This test checks whether past values of one time series (news sentiment) contain information that helps predict another time series (stock prices).

## 5.1 Data Preparation for Granger Causality

The sentiment scores and stock prices were merged by date, ensuring there were no missing values for the analysis.

```python
merged_data = pd.merge(news_data[['publishedAt', 'sentiment']], stock_data[['Date', 'Close']], left_on='publishedAt', right_on='Date')
```

Listing 8: Data Preparation for Granger Causality

## 5.2 Granger Causality Test

Granger causality tests were run for multiple lags to determine if news sentiment "Granger-causes" stock price changes.

```python
from statsmodels.tsa.stattools import grangercausalitytests
granger_test_data = merged_data[['sentiment', 'Close']].dropna()
grangercausalitytests(granger_test_data, maxlag=3, verbose=True)
```

Listing 9: Granger Causality Test

# 6 Results and Discussion

## 6.1 Knowledge Graph Insights

The constructed knowledge graph successfully identified key relationships between entities in the financial news articles. Table 1 presents a sample of the extracted relationships from the dataset. These relationships form the foundation of the knowledge graph, highlighting how different entities such as companies, financial terms, and actions interact in the news.

| **Subject** | **Verb** | **Object** |
|---|---|---|
| LLC | GROW | position |
| LLC | DECREASE | stake |
| Inc. | BOOST | position |
| Inc. | LESSEN | holdings |
| Co. | DECREASE | stake |
| Investments | INCREASE | position |
| Inc. | LESSEN | position |
| that | INCLUDE | suite |
| Co. | GROW | position |
| WA | LIFT | stake |
| LLC | BOOST | holdings |
| LLC | LIFT | holdings |
| LLC | GROW | stake |
| Strategies | DECREASE | stake |

Table 1: Sample of Extracted Relationships from Financial News Articles. Full dataset contains additional relationships.

These relationships were extracted using advanced NLP techniques such as entity extraction and relation extraction, and they serve as the basis for our knowledge graph analysis. The entities and actions reflect trends in how companies and stakeholders adjust their positions in response to market conditions.

## 6.2 Sentiment and Stock Prices

Initial Granger causality results show that there is no significant causal relationship between news sentiment and Apple stock prices within the specified period. The test was conducted with lags of 1, 2, and 3, and in all cases, the p-values were greater than 0.05, indicating a lack of Granger causality. A summary of the test results is provided in Table 2.

Table 2: Granger Causality Test Results

| Lag | Test Name | Test Statistic | p-value | Conclusion |
|---|---|---|---|---|
| 1 | ssr based F-test | 0.1183 | 0.7310 | No significant causality ($p \geq 0.05$) |
| 1 | ssr based Chi2-test | 0.1188 | 0.7303 | No significant causality ($p \geq 0.05$) |
| 1 | Likelihood Ratio Test | 0.1188 | 0.7304 | No significant causality ($p \geq 0.05$) |
| 2 | ssr based F-test | 0.0936 | 0.9107 | No significant causality ($p \geq 0.05$) |
| 2 | ssr based Chi2-test | 0.1885 | 0.9100 | No significant causality ($p \geq 0.05$) |
| 2 | Likelihood Ratio Test | 0.1885 | 0.9101 | No significant causality ($p \geq 0.05$) |
| 3 | ssr based F-test | 0.2384 | 0.8696 | No significant causality ($p \geq 0.05$) |
| 3 | ssr based Chi2-test | 0.7227 | 0.8678 | No significant causality ($p \geq 0.05$) |
| 3 | Likelihood Ratio Test | 0.7223 | 0.8679 | No significant causality ($p \geq 0.05$) |

However, this lack of causality does not rule out the potential for more granular data, such as intraday stock prices or different types of news sources, to provide significant results. Further analysis with such data might yield different conclusions regarding the relationship between sentiment and stock prices.

# 7 Analysis of Existing Issues

## 7.1 Data Sparsity Problem

During the web scraping process, despite setting specific time frames and keywords (e.g., "Apple Inc."), the collected news data exhibited significant sparsity. In particular, during certain date ranges, the quantity of news articles was insufficient, making it difficult to construct a comprehensive financial knowledge graph. This issue could negatively affect subsequent entity recognition, relationship extraction, and the accuracy of causal inference. Moreover, one significant challenge was the limited time frame of the data collection. The financial news articles and stock price data were only collected for the month of July 2024. This short window may not have been sufficient to capture the broader dynamics of how news impacts stock prices. A longer time frame, or more granular intraday stock data, could provide a better basis for analysis and potentially reveal stronger causal relationships.

## 7.2 Simplification of Entity and Relationship Extraction

From the results of entity extraction, most relationships between entities were overly simplified (as shown in Figure 1), mainly consisting of singular entities without complex contextual relationships. This indicates that, during the actual process, the relationship extraction based on the current NLP model yielded overly basic results, failing to capture more complex associations between financial events. This phenomenon reflects the limitations of TextBlob, a relatively simple tool for evaluating the polarity of news content. While TextBlob provides a general sentiment score, it may not be fully optimized for financial texts, where nuances in language can have a large impact on the perceived meaning. Financial jargon and complex phrasing may not be appropriately captured, leading to misclassifications of sentiment.

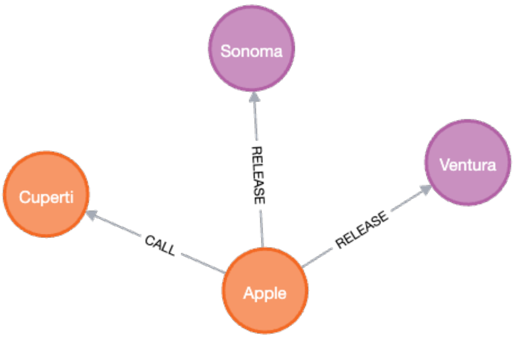

Figure 1: Simplification of entity and relationship extraction

## 7.3 Limitations of Granger Causality Test

After performing Granger causality tests with multiple lags, it was found that the causal relationship between news sentiment and stock closing prices was not significant. This result suggests that the direct causal link between sentiment-based news data and stock fluctuations may be weak. However, this test assumes linear relationships and does not account for the complexities and nonlinearities that often characterize financial markets. Additionally, the test only checks for causality in the time series, which may overlook other influences like economic conditions, investor behavior, or global events that could also affect stock prices.

## 7.4 Simplicity of Results and Visualization Challenges

From the visualization results, although the knowledge graph was successfully generated, the overall relationship graph remains overly simplistic (as shown in Figure 2). The generated graph fails to effectively represent the complex causal relationships and multi-layered financial behavior patterns present in financial news. This could be due to insufficient data, simplistic relationship extraction logic, or the limited applicability of the model to the financial domain. The simplicity of the graph impacts the academic depth of the research and reveals the method's limitations in handling complex financial data.

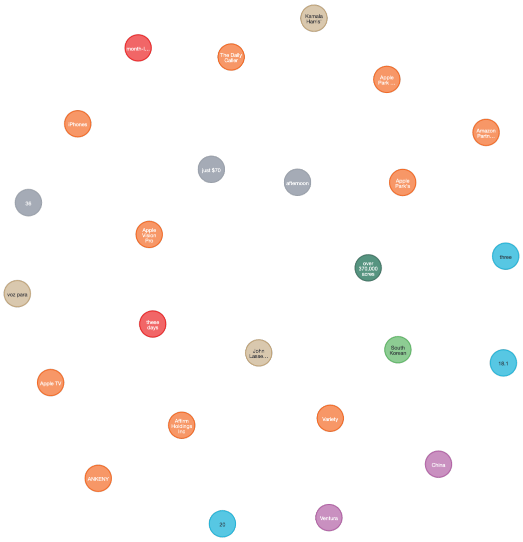

Figure 2: Simplicity of results and visualization challenges