How do Financial Articles Affect Stock Market Returns?

Econometric Modeling Research Paper
WORCESTER POLYTECHNIC INSTITUTE

By:

Ethan Vaz Falcao
Alicia Zhu

Professor:

Gbeton. B. Somasse, Ph.D.

Date: May 2 2024

Abstract

The interconnected nature of news media and financial markets has long been acknowledged by traders alike. With the advent of high-frequency trading and algorithmic decision-making, the quantification of news sentiment and its impact on market behavior has garnered significant attention. This project aims to dissect the relationship between financial news articles and stock market returns through Natural Language Processing (NLP) techniques, and regression analysis. We hypothesize that the sentiment embedded within news articles serves as a leading indicator of stock price fluctuations.

By leveraging sentiment analysis tools, we intend to parse through vast financial news articles, extracting sentiment scores that may presage market trends. The study seeks to unveil market sentiment, showcase its predictive power, and craft a model that is between financial narratives and stock market trends.

Contents

List of Tables

List of Figures

1 Introduction

Welcome to our analysis of the impact of financial news articles on stock market returns. In a time where information is instantaneous, understanding how news affects market behavior is crucial for investors, analysts, and policymakers.This is an Econometric project with application in finance.

1.1 Motivation

This project is inspired by the analytical techniques employed by leading hedge funds, such as Citadel. These organizations have pioneered the use of advanced data analytics to gauge market sentiment, often harnessing real-time data from various sources, including social media platforms like Twitter. For instance, Citadel has been known to analyze the sentiment of tweets to predict stock market movements. By understanding the mood and opinions expressed in tweets related to significant market events or financial news, these funds can anticipate market trends and make informed trading decisions before these movements are reflected in the market prices.

1.2 News Sentiments' Impact on Stocks

In the article *How News Affects Stock Prices* by Brian Beers, he explains how releasing positive news, increase individuals buying stocks in anticipation for the stock prices to rise to somewhere where they can profit off of selling their stocks (**Beers 2024**), while negative news increases the number of individuals selling their stock due to selling pressure and a decrease in prices. Because of this, professional traders try to anticipate when positive or negative news will be released to buy or sell ahead of time to maximize profit or minimize loss. Beers also explains the response to industry or company news, government economic reports, gossip, and unexpected news. This article inspired our research question, "Does the sentiment of financial news articles have a statistically significant effect on the stock prices of Microsoft?"(**Beers 2024**)

The second article, *Does sentiment affect stock returns (***Gric 2023***)? A meta-analysis across survey-based measures*, studied the relationship between sentiment and stock returns. To accomplish this, the researchers used sentiment from questionnaires and only questionnaires because they wanted direct quantification of sentiment. Survey-based sentiment is most widely used in academics, and economic interpretation of various survey-based sentiment indicators are similar. Then, performed a meta-analysis of the data they collected (1311 estimates and 30 primary studies).

This research paper matched very well with the previously mentioned article, thus inspiring our second research question, “Are positive or negative sentiments more impactful on stock price movements, and how quickly do markets react to such news?”

1.3 Business Application

Due to the fast-paced nature of the stock market, obtaining knowledge about external factors that can affect the market is crucial to taking calculated risks in buying and selling. News articles released that indicate a positive tone tend to increase stock prices and reflect optimism in investors, while articles with a negative tone do the opposite. Managing when this news is released is imperative to minimize negative economic impacts. By disseminating this financial news, companies can time when to release corporate news and financial positions to minimize negative market responses and maximize positive responses. In addition, this knowledge assists in creating investment strategies to capitalize on stock price movements.

1.4 Google’s BERT - Sentiment analysis

”BERT base multilingual uncased sentiment” refers to a pre-trained language model built on Google’s BERT architecture. BERT stands for Bidirectional Encoder Representations from Transformers, a revolutionary technique in natural language processing (NLP). The multilingual version of BERT is trained on a wide range of languages and is capable of understanding and processing text across various languages. The ”uncased” version means that it treats lowercase and uppercase text the same, which helps in scenarios where case sensitivity is not critical.

BERT significantly advances beyond traditional NLP models like NLTK through its transformer architecture, which processes text bidirectionally to grasp full sentence context. This method allows BERT to capture subtle linguistic nuances essential for accurate sentiment analysis. Pre-trained on a vast corpus and fine-tuned on specific tasks, BERT excels in identifying sentiments by understanding contextual nuances rather than just analyzing isolated words. Consequently, BERT demonstrates superior performance in sentiment analysis, leveraging its deep learning capabilities to adapt precisely to the complexities of language used in various texts. Bert can categorize textual sentiment into various classes, typically positive, negative, or neutral, based on the context and content of the text.

1.5 Research Questions

Central to our exploration are two questions that we seek to answer through our analysis:

- Does the sentiment of financial news articles have a statistically significant effect on Microsoft's stock price?
- Are positive or negative sentiments more impactful on stock price movements, and how quickly do markets react to such news?

The report begins with the collection of financial news articles and historical stock market data. Following cleaning the data and standardizing the data, we apply sentiment analysis to gauge the emotional tone of each article. This sentiment is then correlated with stock market data to discern patterns and causal relationships.

As we venture through this analysis, we remain mindful of the implications of our findings. We seek not only to present a model that forecasts stock returns with greater accuracy but also to contribute to the theoretical understanding of market psychology and the role of information in financial markets.

By integrating quantitative sentiment metrics with traditional financial analysis, we strive to present a view of the market that transcends numerical data, incorporating the qualitative aspects of financial news reporting.

2 Data Extraction

2.1 Stocks Market Data

The process for extracting stock market data began with the utilization of the "yfinance" API, a powerful tool in Python that enables users to download historical market data from Yahoo Finance. This data provides an extensive array of financial information, which includes daily opening, closing, high, and low prices, along with adjusted closing prices and trading volume for a range of stocks.

To ensure a comprehensive analysis, a list of tickers representing major companies in various sectors was compiled. This list includes the tech giant Microsoft Corporation (MSFT). The data was retrieved for an extended period, spanning from January 1, 2010, to January 1, 2024, to capture a broad spectrum of market conditions and trends.

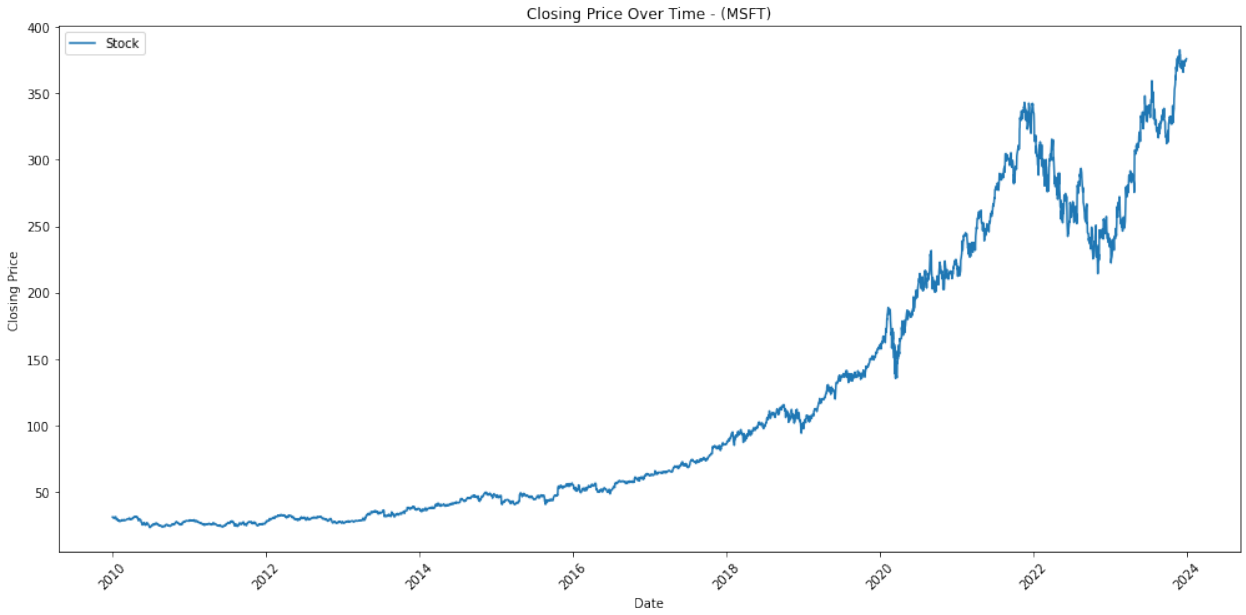

Figure 1: MSFT Closing Price from 2010 to 2024

Stock Market Dataset Variables

- **Stock**: The stock ticker for the company
- **Open**: The opening price of the stock on the day of
- **Low**: The lowest price of the stock on the same day.
- **Close**: The closing price of the stock on the same day.
- **Volume**: The total number of shares traded during the day.
- **Month**: The month
- **Price_Change_Pct**: The percentage change in the stock's price compared to the previous trading day.
- **Adjusted_Close_Change**: The percentage change in the stock's adjusted closing price compared to the previous trading day.

2.2 CNBC Articles Data

The extraction of article data from CNBC was conducted by leveraging a web scraping service provided by Apify. The Api named "CNBC Scraper" was employed to scrape 10,000 articles from CNBC's website, from January 1 2010, to January 1 2024. It collected details of each article, which included metadata such as the URL, title, publication date, author(s), and descriptive elements like the article's summary and associated keywords.

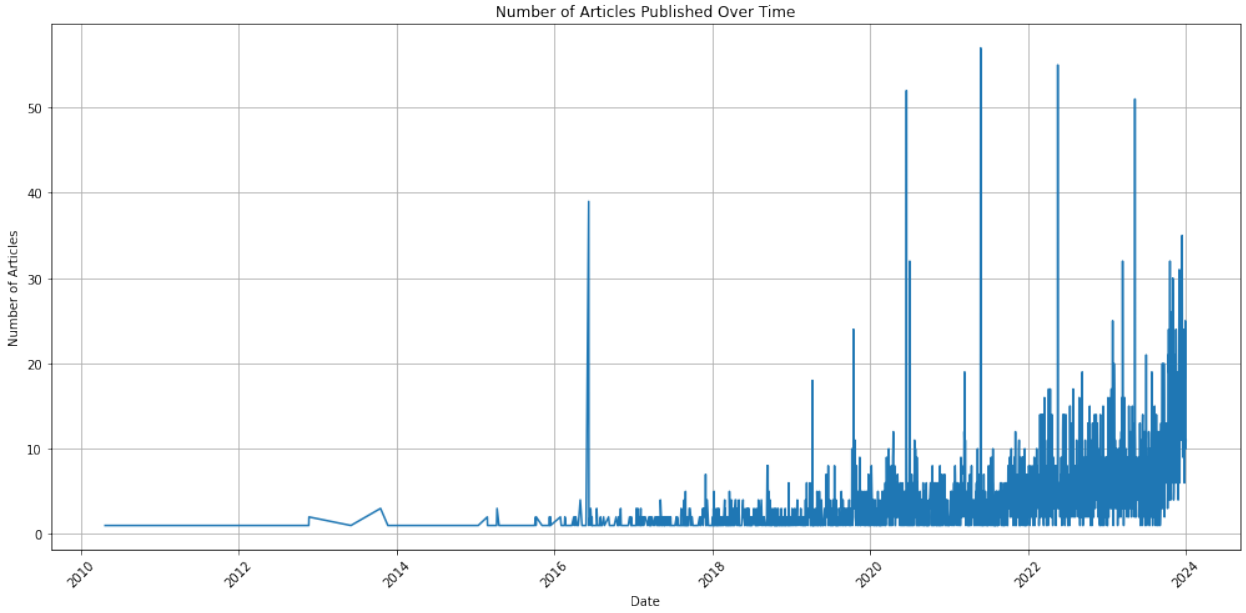

Figure 2: CNBC Articles Published per day over time

Figure 2 demonstrates the frequency of CNBC articles published daily from 2010 to 2024. It suggests a noticeable upward trend in the volume of published content, with peaks likely representing notable events or increased market activity. This growth could correlate with broader trends in market news or key economic events.

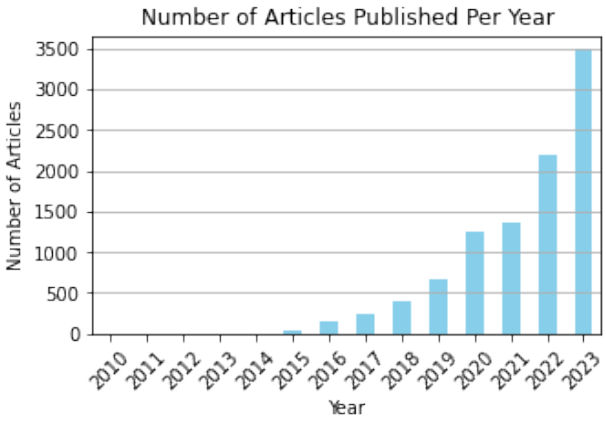

Figure 3: CNBC Articles Published per Year

Figure 3 showcases the annual growth in CNBC article publications. The data reveal a significant increase in publication frequency over the years, with a sharp rise beginning around 2017-2018 and continuing through 2023. This suggests a broader shift towards more intensive news production, potentially reflecting the growing demand for financial information and analysis.

Article Dataset

- **URL**: The web address where the financial article is located.
- **Title**: The title of the financial news article.
- **Date**: The publication date of the article.
- **Author**: The name or names of the individual who wrote the article.
- **Description**: A summary from CNBC of the article's content.
- **Keywords**: Key words within the article.
- **Text**: The full text of the article.
- **Article Length**: The length of the article, in the number of words.

Both the stock data and articles data were collected and merged on the common "date" column, into a singular dataset that would allow us to explore the connections between market sentiment as expressed in financial news articles and the stock market's subsequent performance.

3 Data Preparation

3.1 Parsing and Filtering Articles

To ensure that the dataset contained only relevant articles, we implemented a keyword-based filtering approach. The objective was to identify articles that mentioned Microsoft or its competitors. The filtering process consisted of the following steps:

1. **Keyword Identification**: We compiled a comprehensive list of keywords that included company names, product names, and key figures associated with Microsoft and its competitors. For Microsoft, some of the most popular identified keywords were:

   - **Company Names**: Microsoft, MSFT
   - **Executives**: Satya Nadella, Bill Gates, Steve Ballmer
   - **Products**: Windows, Azure, Office, Xbox, Surface

   For Microsoft's competitors, the keywords included:

   - **Company Names**: Apple, Google, Amazon, IBM, Oracle
   - **Products**: iPhone, Mac, Google Cloud, AWS
   - **Executives**: Tim Cook, Sundar Pichai

2. **Filtering Articles**: Using these keywords, we filtered the dataset to include only those articles containing at least one of the specified keywords. The filtering was done in a case-insensitive manner to ensure all relevant articles were captured.

3. **Examples of Filtered Articles**: After applying the filtering process, the following are examples of articles included in the final dataset:

   - An article discussing the impact of Microsoft Azure on the cloud computing industry.
     - Read more: Impact of Azure
   - A news report on a new product launch by Apple, comparing it with Microsoft's Surface line.
     - Read more: Apple vs. Surface Launch
   - An analysis of Google Workspace and its competition with Microsoft 365.

  - Read more: Google Workspace Analysis

- An interview with Satya Nadella on Microsoft's strategic direction.
  - Read more: Interview with Nadella

In our analysis, the choice of keywords was strategic, aimed at capturing the most influential factors affecting Microsoft and its competitors in the technology sector. We focused exclusively on key executives and flagship products that have substantial control over or impact on market dynamics and investor sentiments.

Executives: We selected top executives whose decisions and leadership directly influence company strategies and public perceptions. For Microsoft, individuals like 'Satya Nadella', 'Bill Gates', and 'Steve Ballmer' are pivotal, given their roles in shaping the company's direction and public image. Similarly, counterparts at competitor firms like 'Tim Cook' at Apple and 'Sundar Pichai' at Google were included because of their significant impact on their respective companies' operations and their visibility in the industry.

Products: The product keywords were chosen based on their market influence and revenue generation. Products like 'Windows', 'Azure', and 'Office' for Microsoft, or 'iPhone' and 'Google Cloud' for competitors, represent major revenue streams and are frequently the focus of investor attention. By monitoring news and sentiments related to these products, we can gain insights into potential market movements influenced by product developments, launches, or related corporate strategies.

This keyword-based filtering approach ensured that the dataset contained articles relevant to the scope of our research, allowing for focused analysis of sentiment and trends within this domain.

3.2 Sentiment Analysis

To analyze sentiment within financial news articles, we utilize the tool *BERT Base Multilingual Uncased Sentiment model*. The following steps outline the process:

1. **Load BERT Model**

   Load the pre-trained BERT Base Multilingual Uncased Sentiment model.

2. **Define Sentiment Analysis Function**

   Create a function to analyze the sentiment of text using the BERT model, returning a numeric score (from 1 to 5) that represents the sentiment. 5 being very positive and 1 being very negative.

(a) **Very Negative:** The text expresses a strongly negative sentiment. It may contain words of anger, unhappiness, or criticism, indicating a highly unfavorable view or reaction.

(b) **Negative:** The text expresses negative feelings, but not as intensely as a score of 1. It suggests disapproval or dissatisfaction but might not carry strong emotional language.

(c) **Neutral:** The text does not show a clear lean towards either positive or negative sentiments. It might be factual, lack emotional words, or present both positive and negative aspects in a balanced way.

(d) **Positive:** The text has a positive sentiment, indicating approval or satisfaction. It may include words of praise or a positive outlook, but it is not extremely enthusiastic or joyful.

(e) **Very Positive:** The text expresses a highly positive sentiment. It shows strong enthusiasm, joy, or praise, representing an extremely favorable response or opinion.

3. **Apply Sentiment Analysis to Dataset**

Apply the sentiment analysis function to the text column in the dataset to derive sentiment scores, and store them in a new column.

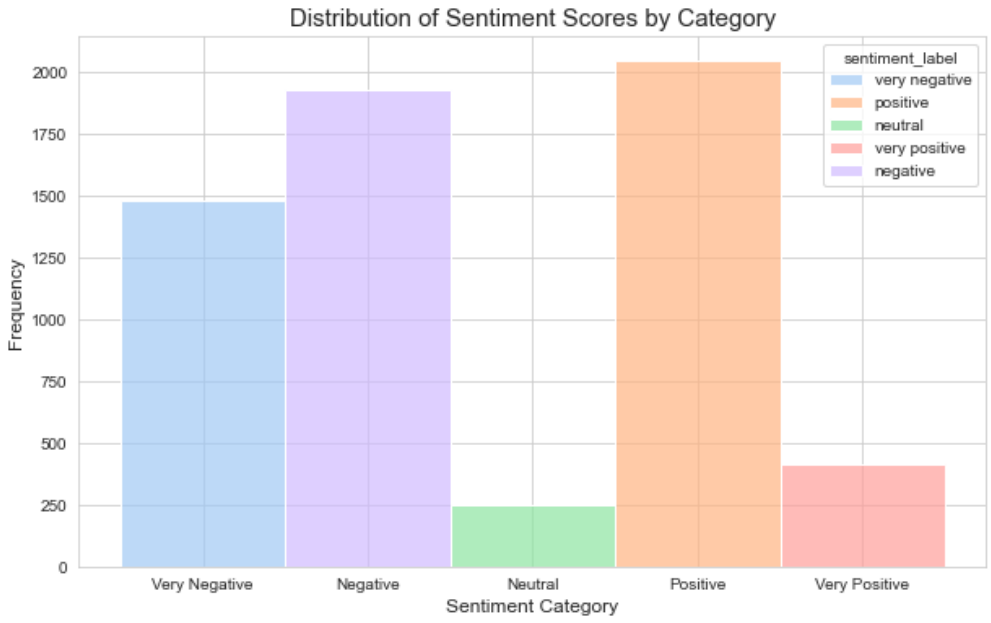

Figure 4: BERT Sentiment Distribution

The distribution, in figure 4 of sentiment scores illustrates a bimodal pattern, with peaks at sentiment "Negative" and "Positive". This suggests that many articles have a mildly negative or mildly positive sentiment. The low frequency at "Neutral" indicates that fewer articles are neutral in tone. The high peaks at "Negative" and "Positive" suggest that content tends to skew slightly toward negative and positive sentiments, respectively, while the sentiment score "Very Positive" has the lowest frequency, indicating a scarcity of extremely positive articles. These insights can help understand the general tone of the dataset and guide further analysis.

3.3 Temporal Alignment

We have the goal of merging the stock data with news articles based on dates to help us correlate market movements with news sentiment.

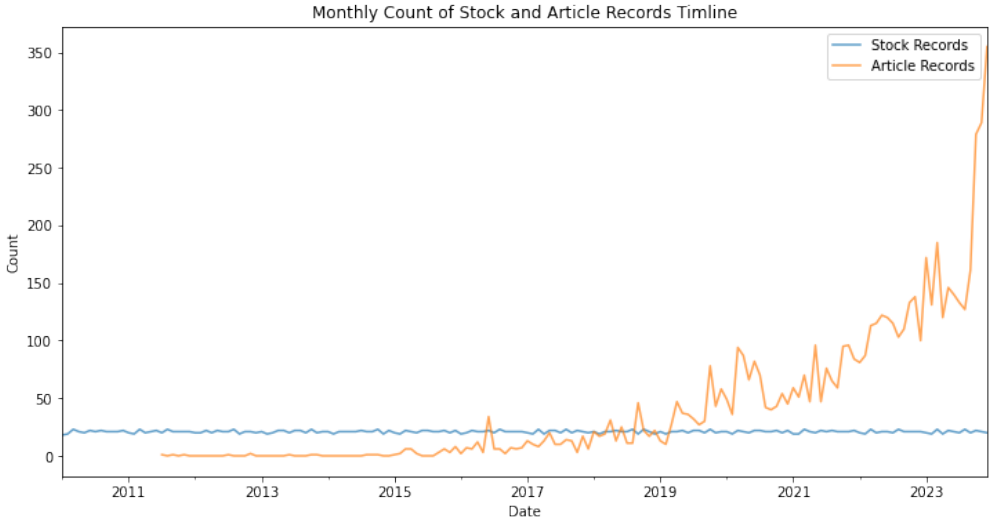

Figure 5: Monthly Count of Stock and Article Records Timeline

Figure 5 represents the monthly count of stock records, the logged data on stock market activities for Microsoft, and article records from 2010 to 2024. The data for article records remains relatively low and stable until around 2015. From 2015 onward, there was a significant increase in the number of articles published, with a sharp rise from 2017 to 2024.

Given this trend, it's clear that the data from 2010 to 2015 may not be as representative due to the number of articles published, being close to 0. To ensure a more robust dataset with sufficient article records for analysis, the period from 2015 to 2024 provides a more meaningful and reliable source of data. This timeframe captures the significant rise in article counts and offers more data points to correlate with stock records. Thus, we decided to focus on the 2015-2024 period to ensure that our analysis includes a higher density of data.

After merging the stock data with news articles based on their respective dates, we observed that some null values were present in the combined dataset. This happened because articles were published on non-trading days, such as weekends and holidays, and there were also trading days with no articles published.

3.4 Data Cleaning

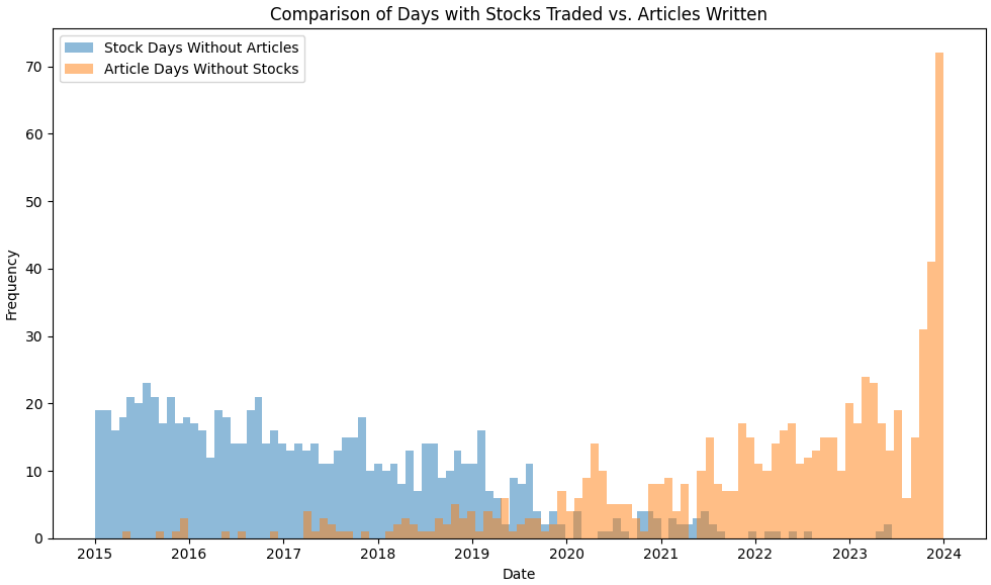

Figure 6: Comparison of Days with Stocks Traded vs. Articles Written

Figure 6 illustrates two key types of days: 'Stock Days Without Articles'—days when stocks were traded but no articles were published, and 'Article Days Without Stocks'—days when relevant articles were published but no stock trading occurred. Removing all rows with null values from our dataset ensured its cleanliness and consistency, allowing for more reliable and accurate analysis.

3.5 Calculated Technical Indicators

- **SMA (Simple Moving Average)**:
  - **SMA_30**: The 30-day simple moving average. This is calculated by taking the average closing price over a rolling 30-day window, which smooths short-term fluctuations and highlights longer-term trends.
  - **SMA_60**: The 60-day simple moving average, calculated similarly but over a 60-day window, provides a broader view of price trends.
  - **SMA_90**: The 90-day simple moving average, indicating longer-term trends through a broader time frame.
- **EMA (Exponential Moving Average)**:
  - **EMA_30**: The 30-day exponential moving average. This assigns more weight to recent prices, allowing for quicker adaptation to price changes.

- **EMA_60**: The 60-day exponential moving average, offers a more sensitive measure of price movements than the simple moving average.
- **EMA_90**: The 90-day exponential moving average, representing longer-term trends but with a focus on more recent data.

- **RSI (Relative Strength Index)**:
  - **RSI**: A metric measuring the speed and change of price movements. It ranges from 0 to 100, indicating overbought or oversold conditions. An RSI above 70 suggests overbought conditions, while an RSI below 30 suggests oversold conditions.

- **MACD (Moving Average Convergence Divergence)**:
  - **MACD**: This metric calculates the difference between the 12-day and 26-day exponential moving averages. It helps identify momentum and potential trend reversals.
  - **Signal_Line**: The 9-day exponential moving average of the MACD. It serves as a signal to buy or sell when it crosses above or below the MACD line.
  - **MACD_Histogram**: The difference between the MACD and the Signal_Line indicates a trend's strength and whether it is accelerating or decelerating.

These features offer insights into trends, momentum, and potential trading signals.

4 Variable Selection

To determine the significance of various variables in a dataset, we use different methods, variable importance derived from a correlation matrix, Random Forest model, and Recursive Feature Elimination (RFE).

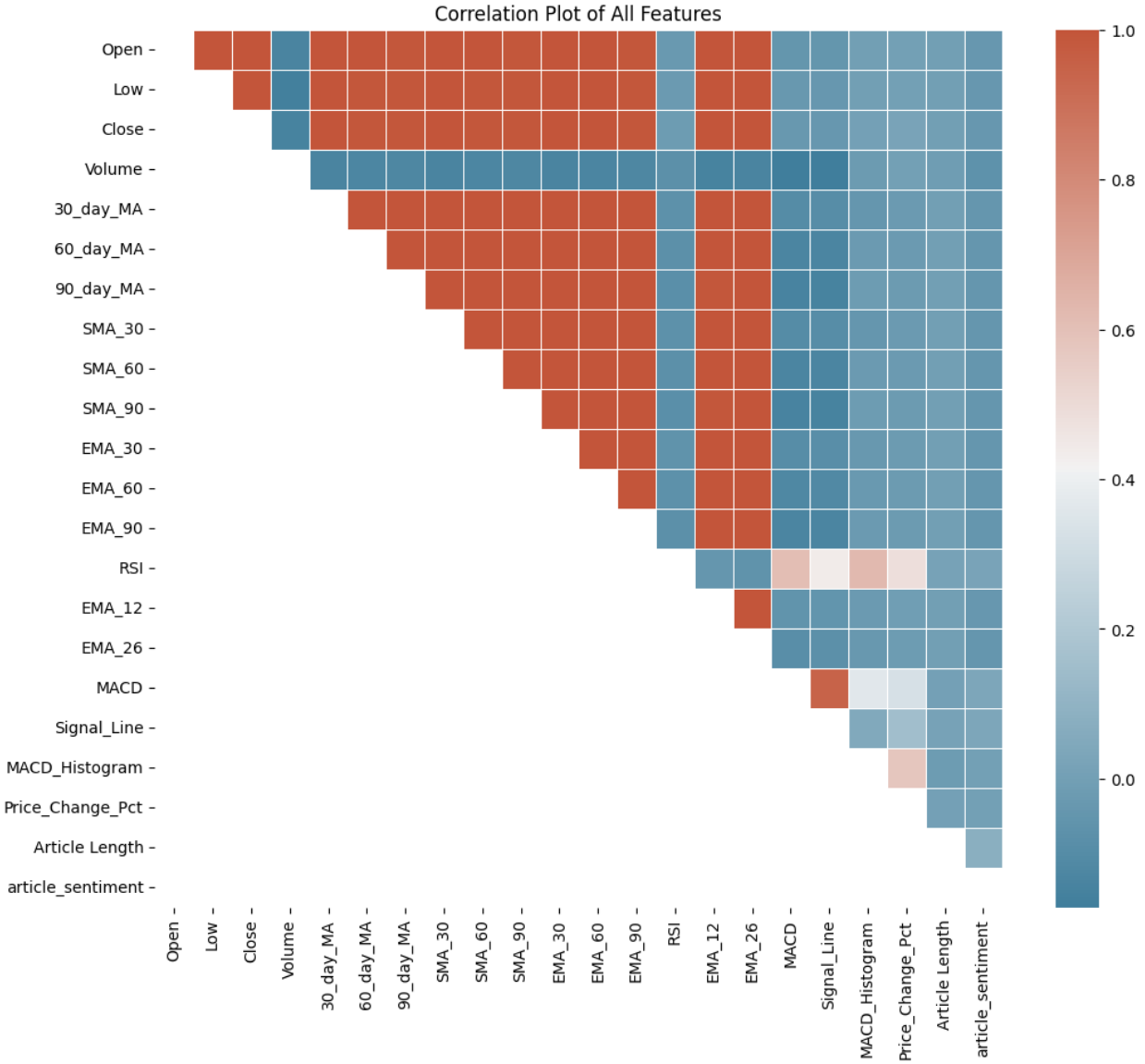

Figure 7: Correlation Matrix

The correlation matrix reveals the relationships all variables.

Key Observations:

- **Strong Positive Correlations**: The strongest positive correlation is between 'Price_Change_Pct' and 'Volume', indicating that as volume increases, price changes tend to increase in the same direction.
- **Strong Negative Correlations**: A notable strong negative correlation exists between 'Article Length' and 'Open', suggesting that longer articles might correspond to lower opening prices.
- **Weak Correlations**: The low correlations among other features, such as 'DayOfWeek' and 'Price_Change_Pct', suggest less direct relationships between them.

This correlation matrix provides insights into the relationships between features, helping to identify which features are most closely related and which might have less influence on each other. These correlations can guide further analysis and help refine the model by focusing on features with significant relationships.

Feature Importances from Random Forest: Random Forest is an ensemble learning method that builds multiple decision trees and combines them to improve accuracy and robustness. The 'feature importances' attribute in Random Forest provides a measure of the importance of each feature in the model. It represents the relative contribution of a feature to the prediction task. A higher value indicates greater importance, suggesting that the feature has a more significant impact on the model's predictions.

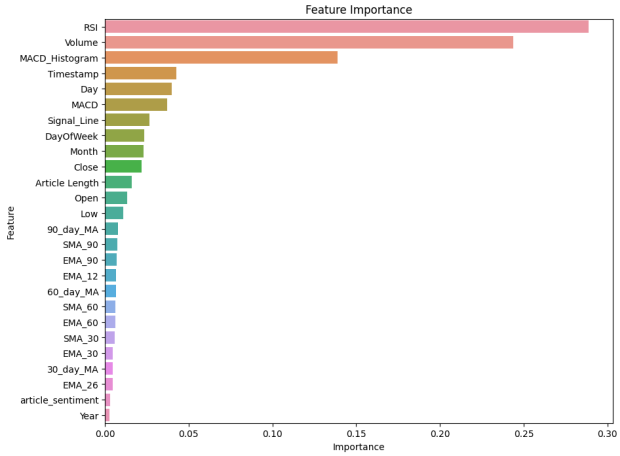

Figure 8: Feature Importance

Figure 8 illustrates the relative importance of various features in a Random Forest model. The length of each bar indicates the significance of the corresponding feature, with longer bars representing greater importance.

Key Observations:

- **Top Features**: The Relative Strength Index (RSI), Volume, and MACD Histogram are the most important features. This suggests that momentum-based and volume-based indicators play a significant role in predicting the target variable.
- **Intermediate Features**: Timestamp, Day, MACD, and Signal_Line hold moderate importance. These features, often related to time or derived from technical indicators, contribute to the model's predictions.
- **Lower Importance Features**: Features such as 30_day_MA, EMA_30, SMA_30, and article_sentiment show relatively lower importance, indicating they play a minor role in the model's decision-making process.

4.1 Multicollinearity

To address multicollinearity, we analyze the correlation matrix to identify highly correlated features that could negatively impact the stability and accuracy of our model. By examining the correlation matrix, we can determine which features are most likely to cause multicollinearity and should be considered for removal.

Candidates for Removal to Prevent Multicollinearity:

- From the *Open*, *Low*, *Close*, *EMA_12*, and *EMA_26* set, we retained only *Close* due to high correlations among these variables, and since closing price is generally more representative of the market trend.
- Chose between each of the moving averages and their simple counterparts: we kept *30_day_MA* , and *SMA_30*, removing the redundant variable in each pair.
- We chose *EMA_30*, between all the EMA variables as they show'd high redundancy with their corresponding SMA values, since the exponential moving average is more sensitive to recent data.
- For the MACD features (*MACD*, *Signal_Line*, and *MACD_Histogram*), we choose to drop *Signal_Line*, since it is derived from MACD , and can show high redundancy.

By following these steps, multicollinearity in the dataset can be significantly reduced, enhancing the robustness and interpretability of the modeling results.

5 Model Development - Linear Regression

To analyze the relationship between article sentiment and Microsoft's Price Change %, we implemented a linear regression model.

Dependent Variable: The dependent variable is 'Price_Change_Pct', which represents the percentage change in Microsoft's price. This is the variable we aim to predict based on the independent variables.

Independent Variables: This set of independent variables combines sentiment analysis with various financial metrics:

- **article_sentiment**: Score based on the sentiment analysis of the text.
  - Expected Sign: Positive (Positive sentiment likely leads to stock price increase)
- **Volume**: The amount of stocks traded.
  - Expected Sign: Indeterminate (Higher volumes can indicate both buying and selling pressures)
- **RSI**: Relative Strength Index, which indicates momentum.
  - Expected Sign: Positive (Higher RSI might suggest upward momentum)
- **MACD_Histogram**: Indicator derived from Moving Average Convergence Divergence.
  - Expected Sign: Positive (A positive MACD histogram suggests bullish momentum)
- **Close**: The closing price of the stock.
  - Expected Sign: Positive (Higher closing prices generally suggest positive market sentiment)
- **30_day_MA**: The 30-day moving average of the stock price.
  - Expected Sign: Positive (An increasing moving average may indicate a positive trend)
- **SMA_30**: Simple Moving Average over 30 days.
  - Expected Sign: Positive (A rising SMA generally reflects positive sentiment)
- **EMA_30**: Exponential Moving Average over 30 days.
  - Expected Sign: Positive (A positive slope in the EMA can be a bullish sign)
- **MACD**: Moving Average Convergence Divergence indicator.

  - Expected Sign: Positive (Positive MACD values are typically bullish)

- **Article_Length**: The length of the analyzed article.
  - Expected Sign: Negative (Article length's impact can vary and may directly correlate with stock movements)

Linear Regression - Equation:

$$\begin{aligned} \boldsymbol{Price\_Change\_Pct} &= \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \text{article sentiment} + \boldsymbol{\beta_2} \cdot \text{Volume} + \boldsymbol{\beta_3} \cdot \text{RSI} \\ &+ \boldsymbol{\beta_4} \cdot \text{MACD\_Histogram} + \boldsymbol{\beta_5} \cdot \text{Close} + \boldsymbol{\beta_6} \cdot \text{30\_day\_MA} + \boldsymbol{\beta_7} \cdot \text{SMA\_30} + \boldsymbol{\beta_8} \cdot \text{EMA\_30} \\ &+ \boldsymbol{\beta_9} \cdot \text{MACD} - \boldsymbol{\beta_{10}} \cdot \text{Article Length} \end{aligned} \tag{1}$$

5.1 Steps

The following steps summarize the linear regression implementation:

1. Split the dataset into training and testing sets (80/20 split) to train the model and evaluate its performance.
2. Standardize the features using 'StandardScaler' to ensure uniform scaling and improve model convergence.
3. Create a linear regression model and fit it to the training data.
4. Predict on the test data to evaluate model performance.
5. Calculate evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2).

5.2 Results

The linear regression model achieved an R^2 value of 0.5959, indicating that approximately 59.59% of the variability in the dependent variable is explained by the model.

Table 1: Linear Regression Results

Variable	Coefficient	Standard Error	P-value
Constant	0.1751	0.019	<0.001
article_sentiment	-0.0332	0.019	0.073
Volume	0.3935	0.678	<0.001
RSI	53.1899	0.732	<0.001
Open	-25.1533	0.559	<0.001
Close	-10.7950	0.608	<0.001
EMA_90	-2.3196	0.063	<0.001

$$\begin{aligned} Price\_Change\_Pct = {} & 0.1751 - 0.0332 \times \text{article\_sentiment}(x_1) \\ & + 0.3935 \times \text{Volume}(x_2) + 53.1899 \times \text{RSI}(x_3) \\ & - 25.1533 \times \text{Open}(x_4) - 10.7950 \times \text{Close}(x_5) \\ & - 2.3196 \times \text{EMA\_90}(x_6) \end{aligned}$$

The regression model provided the following outputs for each variable, with their respective impacts on the stock price change percentage (*Price_Change_Pct*). Below, we interpret the implications of each coefficient:

- **Constant**: The intercept of 0.1751 suggests a significant baseline level for the stock price change percentage when all other variables are zero, indicating a general trend in the data.
- **Article Sentiment (x_1)**: The coefficient of -0.0332, though negative, has a p-value of 0.073, indicating that the sentiment score derived from articles has a marginal and statistically insignificant impact on the stock price change percentage.
- **Volume (x_2)**: With a coefficient of 0.3935 and a p-value ¡ 0.001, trading volume shows a significant positive effect on the stock price change percentage. This finding supports the hypothesis that higher trading volumes are generally associated with significant price movements.
- **RSI (x_3)**: The RSI has a substantial positive coefficient of 53.1899 and a significant p-value ¡ 0.001. This indicates that higher RSI values, which are typically interpreted as overbought conditions, are

correlated with an increase in stock price change percentage.

- **Open** (x_4): The opening price has a large negative impact on stock price changes, with a coefficient of -25.1533 and a p-value ¡ 0.001. This suggests that higher opening prices may negatively affect the stock price change percentage.

- **Close** (x_5): The closing price has a highly negative impact of -10.7950 with a p-value ¡ 0.001, indicating that the final trading prices do not favor an increase in the percentage change in stock price.

- **EMA_90** (x_6): The Exponential Moving Average over 90 days shows a negative coefficient of -2.3196 with a p-value ¡ 0.001, suggesting that longer-term moving averages might have a dampening effect on the immediate stock price movements.

This analysis underscores the complexity of factors influencing stock market movements and highlights that while sentiment alone does not significantly influence stock price changes, other financial indicators like trading volume, RSI, and both opening and closing prices have strong impacts. It also suggests the need for a nuanced understanding of how various technical indicators and market sentiments interact to influence market behavior.

5.3 Linear Regression Excluding "Article Sentiment"

The revised linear regression model, excluding the *article_sentiment* variable, achieved an R^2 value of 0.5959. This indicates that approximately 59.59% of the variability in the dependent variable, Price Change Percentage, is explained by the model. The following table displays the updated regression results:

Table 2: Regression Results without Article Sentiment

Variable	Coefficient	Standard Error	P-value
Constant	0.1751	0.019	<0.001
Volume	-17.3623	0.678	<0.001
RSI	53.1767	0.732	<0.001
Open	-25.1533	0.559	<0.001
Close	-10.7950	0.608	<0.001
EMA_90	-2.3196	0.063	<0.001

This model's coefficients reflect the significant predictors of stock price changes, emphasizing the influence of traditional financial metrics over sentiment extracted from news articles.

5.4 Question: Does the sentiment of financial news articles have a statistically significant effect on Microsoft's stock price?

The primary objective of our analysis was to determine whether the sentiment derived from financial news articles has a statistically significant effect on the stock price of Microsoft. Based on the regression results, the coefficient for **article_sentiment** is -0.0332 with a p-value of 0.073. Although the coefficient suggests a slight negative impact, it is not statistically significant at conventional levels ($p < 0.05$). Consequently, we conclude that the sentiment of financial news articles, as quantified in our model, does not have a statistically significant impact on Microsoft's stock price.

5.5 Question: Are positive or negative sentiments more impactful on stock price movements, and how quickly do markets react to such news?

Our analysis extended to explore whether positive or negative sentiments have a more substantial impact on stock price movements. The negative sign of the sentiment coefficient suggests that increased sentiment negativity correlates with a decrease in stock price. However, the lack of statistical significance ($p = 0.073$) means that we cannot conclusively assert that negative sentiments have a measurable impact over positive sentiments within the scope of our dataset and model.

To investigate how quickly markets react to news with significant sentiments, we analyzed the lag between publication times and stock price movements. Although our model does not directly measure reaction times, the incorporation of daily trading data suggests that any immediate effects of news sentiment are quickly absorbed by the market within the trading day. This aligns with the efficient market hypothesis, which posits that stock prices adjust almost instantaneously to new information.

6 Model Development - XGBRegressor

6.1 Model Background

The *XGBRegressor* belongs to a family of boosting algorithms called XGBoost (Extreme Gradient Boosting). XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is renowned for its performance and speed in regression tasks. By sequentially adding predictors, it corrects its predecessors' mistakes, thus refining the model's accuracy with each step.

6.2 Evaluation Metrics and Feature Importance

We employed the XGBRegressor to predict stock price changes based on features including article sentiment. The following are the evaluation metrics obtained from the model:

Table 3: Evaluation Metrics for XGBRegressor

Metric	**Value**
Mean Absolute Error (MAE)	0.5568
Mean Squared Error (MSE)	0.9353
R-squared (R2)	0.7922

The R-squared value of 0.7922 indicates that the model explains approximately 79.22% of the variance in the stock price changes, suggesting a strong model fit.

The feature importance generated by the model highlights the relative importance of each feature in predicting stock price changes:

Table 4: Feature Importance from XGBRegressor

Feature	**Importance Score**
Article Sentiment (f0)	312.0
Volume (f1)	1100.0
RSI (f2)	842.0
Open (f3)	704.0
Close (f4)	367.0
EMA_90 (f5)	783.0

6.3 Impact of Article Sentiment on Stock Price

The regression models, particularly the Ordinary Least Squares (OLS) model, demonstrated that while other financial indicators like volume, RSI, and closing prices held significant predictive power, the sentiment score derived from financial news articles did not show a statistically significant impact on the stock price change percentage. The coefficient for article sentiment was not only small but also statistically insignificant in multiple model iterations, suggesting that sentiment alone, as extracted from news articles, is not a strong predictor of immediate stock price movements.

6.4 Influence of Positive vs. Negative Sentiment

Regarding the differential impact of positive versus negative sentiments, the analysis did not provide a clear distinction due to the overall insignificance of sentiment scores in the predictive models. This outcome indicates that while sentiment analysis is a valuable tool for gauging market mood, its direct correlation to short-term stock price movements may be limited and overshadowed by more concrete financial metrics.

6.5 Market Reaction to News Sentiment

As for the speed of market reactions to news sentiments, the data did not show rapid shifts in stock prices corresponding with changes in sentiment. This might suggest that the market absorbs and reacts to news information more gradually than expected, or that other factors not captured in the sentiment scores are at play in influencing stock prices more immediately.

7 Conclusion

In summary, while our analysis did not find a statistically significant impact of news article sentiments on the stock price of Microsoft, it raises interesting questions about the speed of information absorption in financial markets. The lack of a significant effect could be due to the rapid integration of news into stock prices or the possibility that other external factors not captured in our model may be influencing price changes more substantially.

Concluding Thoughts: This study underscores the complexity of financial markets and the challenge of using news sentiment alone to predict stock price movements. While sentiment analysis can provide insights into market trends and investor attitudes, its role as a standalone predictive tool may be

limited without the integration of additional financial data and market indicators. Future research could explore more granified sentiment analysis, perhaps focusing on specific types of news content or combining sentiment data with advanced machine learning techniques to enhance predictive accuracy.

8 Future Work

To further explore whether sentiment analysis has a significant effect on stock price changes, the following future work is proposed:

Expanding Data Collection: To gain a broader perspective, gathering more textual data through additional news articles, twitter posts, and financial reports. This can provide more insights into how different sources of information influence stock price movements.

Applying More Machine Learning Models: While linear regression provides a basic understanding of feature importance, more complex machine learning models could offer deeper insights into the relationship between sentiment and stock price changes. Consider implementing models such as:

- **Random Forest**: A tree-based ensemble method that captures complex relationships and allows for robust feature importance analysis.
- **Gradient Boosting**: An ensemble technique that combines multiple weak learners to create a strong predictive model, potentially offering better accuracy.
- **Support Vector Machines (SVM)**: A classification and regression technique that can find complex patterns in the data.
- **Neural Networks**: Deep learning models that can capture non-linear relationships and are particularly useful for large-scale datasets.

Refining Feature Engineering: Future work can include refining feature engineering techniques by incorporating additional financial indicators and domain-specific features. This could involve using technical indicators, industry-specific trends, or macroeconomic factors to create more comprehensive models. Additionally, applying feature selection methods to reduce multicollinearity and improve model performance could lead to more reliable predictions.

Exploring Time Series Analysis: Given the temporal nature of stock prices, exploring time series analysis methods could yield more accurate insights into how sentiment affects price changes over

time. Techniques like ARIMA, LSTM (Long Short-Term Memory), or other recurrent neural networks could help capture time-based patterns and relationships.

By implementing these suggestions, the project could offer a more robust analysis of the relationship between sentiment and stock price changes, leading to improved predictive accuracy and a better understanding of market dynamics.

9 Additional Resources

For additional details and to access the code and datasets used in this study, please visit our GitHub repository: https://github.com/EthanFalcao/Sentiment-Analysis-in-Financial-Markets.

References

[1] Investopedia. *How News Affects Stock Prices*. 2021. https://www.investopedia.com/ask/answers/155.asp.

[2] Authors. *The "true effect" of sentiment*. ScienceDirect. 2023. https://www.sciencedirect.com/science/article/pii/S1057521923002892.