Preliminary Data Exploration

- **Features explored**
    - Stock opening price S_d^{open} on a given versus the stock news sentiment scores s_d^{stock} and general news sentiments s_d^{news} on a given day d .
    - Stock closing price S_d^{close} versus the sentiment scores s_d.
    - Difference between the opening and closing $\Delta S_d = S_d^{closing} - S_d^{open}$ prices against the daily stock news sentiment scores s_d^{stock} and general news sentiments s_d^{news} on a given day d .
    - Difference between the opening on day d and closing on day $d + 1$ $\Delta S'_d = S_{d+1}^{close} - S_d^{open}$ prices against the daily sentiment scores s_d.
    - Binary stock price variable $I_{d,d-1} = 1 \; if \; \Delta S'_d > 0$ and $I_{d,d-1} = 0 \; if \; \Delta S'_d < 0$.
- **Models explored**
    - Simple linear regression SLR for ΔS_d versus s_d (both general and stock news).
    - KNN regression for ΔS_d versus s_d (both general and stock news).
    - Logistic regression for the binary stock price variable against the sentiment score s_d^{stock}.
- **Preliminary Results**
    - *SLR model*

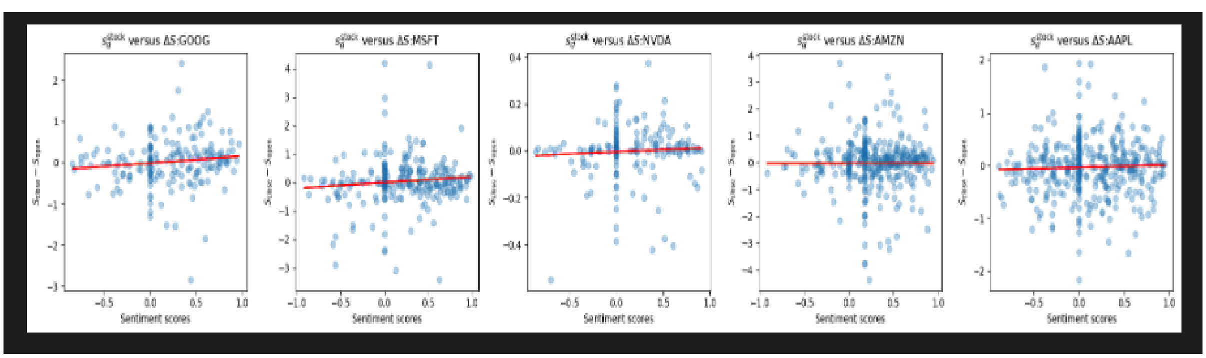


Figure 1. Scatter plots for each stock symbol show sentiment scores vs. stock price difference, with red linear regression lines. The mostly flat lines indicate weak but positive correlation across the stocks.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.014
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     2.683
Date:                Sun, 03 Nov 2024   Prob (F-statistic):              0.103
Time:                        00:08:56   Log-Likelihood:                -160.41
No. Observations:                 185   AIC:                             324.8
Df Residuals:                     183   BIC:                             331.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0108      0.047     -0.229      0.819      -0.104       0.082
x1             0.1656      0.101      1.638      0.103      -0.034       0.365
==============================================================================
Omnibus:                       35.325   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              189.554
Skew:                          -0.512   Prob(JB):                     6.90e-42
Kurtosis:                       7.852   Cond. No.                         2.49
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Confidence intervals:
 [[-0.10403853  0.08237083]
 [-0.03387231  0.36516854]]
```

Figure 2. This table shows OLS regression results for GOOG, with sentiment scores as the independent variable. The low R-squared (0.014) and non-significant slope suggest a weak relationship between sentiment and stock price difference.

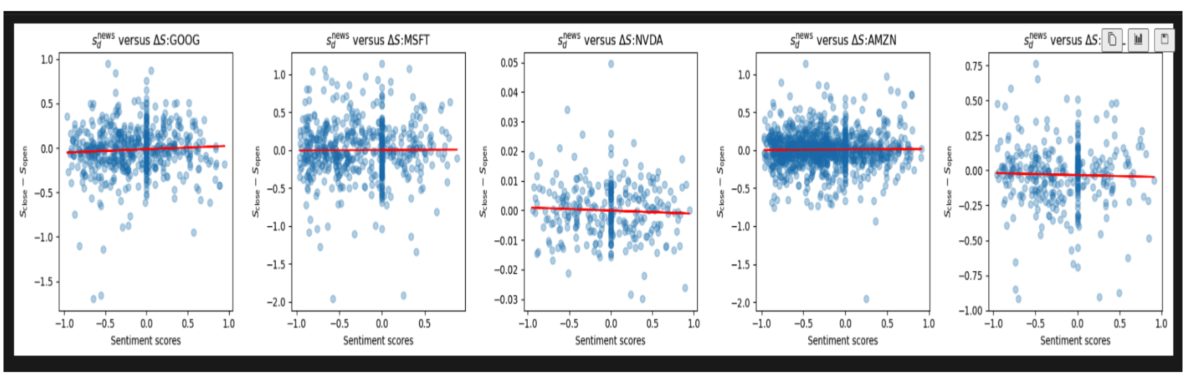


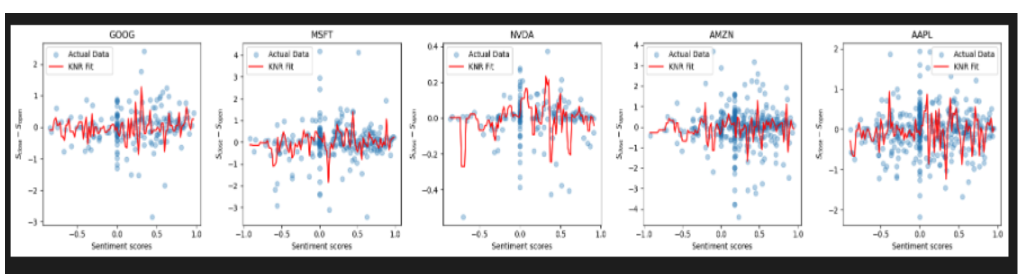
Figure 3. This figure shows scatter plots and linear regression lines (in red) for sentiment scores versus daily stock price differences for GOOG, MSFT, NVDA, AMZN, and AAPL. The nearly flat lines suggest minimal correlation between sentiment scores and stock price changes.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.012
Date:                Sun, 03 Nov 2024   Prob (F-statistic):              0.157
Time:                        19:20:05   Log-Likelihood:                -56.768
No. Observations:                 604   AIC:                             117.5
Df Residuals:                     602   BIC:                             126.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0141      0.012     -1.229      0.220      -0.037       0.008
x1             0.0379      0.027      1.418      0.157      -0.015       0.090
==============================================================================
Omnibus:                      150.777   Durbin-Watson:                   2.080
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              964.820
Skew:                          -0.938   Prob(JB):                    3.11e-210
Kurtosis:                       8.901   Cond. No.                         2.53
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Confidence intervals:
 [[-0.03674896  0.00846068]
 [-0.01458536  0.09044089]]
```

Figure 4. OLS regression for GOOG stock also shows a low R-squared, with no significant relationship between sentiment and the stock price difference. But a p-value of 22% means we are unable to rule out the null hypothesis that there is no relationship between the two.

- **<u>On going work</u>**
  - *Checking for non-linearity via parametric bootstrap*
    - We have a kNN regression fit for our data which has a far smaller MSE than the linear regression fit



    - Our current goal is to assess the statistical significance of the fit and to identify any potential non-linearity in the data. To achieve this, we're currently using parametric bootstrapping, which involves generating resampled datasets based on estimated model parameters. This approach will allow us to evaluate how frequently similar or more extreme results would be expected under the null hypothesis, helping to determine if the observed fit is statistically significant and if there are any signs of non-linearity in the relationship.

- *Exploring timeline for the predictors*
  - We want to investigate the appropriate timeline to compare the stock price with respect to the sentiment scores.
  - For example, perhaps the average stock price over a timeline T is more significantly correlated to the sentiment scores than the daily stock price data. Or the difference between the opening price on day d and the closing price on $d + T$ is a better parameter to predict against the sentiment scores (again averaged over the timeline T).

- *Correlation between various features/predictors*

Stocks Data	**News Data**
Opening	Compound stock news sentiment score
Closing	Compound general news sentiment score
Volume	Positive stock news sentiment score
High	Positive general news sentiment score
Low	Negative stock news sentiment score
Volume	Negative general news sentiment score
Dividends	Neutral stock news sentiment score
Stock Splits	Neutral general news sentiment score

We are trying to obtain a correlation matrix between these features which would enable us to conduct a more robust features selection analysis.

- *Binary stocks function as a response variable*
  - We have used Logistic regression as a model to predict $I_{d,d-1}$ from the sentiment scores and achieved an accuracy of only about 50%.
  - In order to do better we now aim to define a more general version of this function over a certain timeline $I_{d,d-T}$ and then test its behavior against the sentiment score via the Logistic regression model.

- *Time series analysis of the stocks and news data*
  - We now aim to explore the time series based models like **ARIMA** and **SARIMA** and study the corresponding autocorrelation functions.