CIS 9665 Applied Natural Language Processing
Team 5

Kunmei Huang, Suin Kim, Nankun Liang, Swan Htet Linn, Angad Minhas

Table of Contents

Introduction

This study investigates the potential relationship between news headlines and stock market trends, aiming to determine whether sentiment and content within headlines can predict market movements. Using a comprehensive dataset of financial news and corresponding market indices, natural language processing (NLP) techniques are applied to quantify sentiment and extract key themes. Statistical analysis and machine learning models were employed to evaluate the predictive power of headline sentiment on stock price trends. Preliminary results indicate a significant correlation between news sentiment and short-term market fluctuations, offering valuable insights for investors and analysts seeking to integrate news data into trading strategies.

Motivations of Research Question

The Efficient Market Hypothesis (EMH) asserts that stock prices fully incorporate all available information at any point in time. Under this theory, new information—such as economic reports, corporate earnings, or geopolitical events—is immediately reflected in stock prices, assuming rational investors have equal access to information (Fama, 1997). Traditional financial analysis often emphasizes structured numerical data, such as historical prices or financial ratios. However, unstructured news text contains valuable contextual information, including sentiment, tone, emerging developments, and implicit signals, which may not be captured through quantitative indicators alone. The advancements in Natural Language Processing, particularly tools like sentiment analysis and word importance techniques, have created opportunities to analyze textual data and identify patterns or signals relevant to financial markets.

This research explores whether unstructured financial news holds predictive power over stock price movement. Specifically, the investigation focuses on the following questions:

1. Can textual features—such as sentiment, word importance, and Named Entity Recognition —effectively predict whether the Dow Jones Industrial Average (DJIA) will rise or fall?
2. What types of news and specific keywords influence stock prices, and how do these factors manifest in price changes?

Uncovering meaningful relationships between news content and stock price behavior has the potential to provide valuable insights for refining trading strategies, enhancing market analysis, and deepening the understanding of how unstructured information drives financial decision-making.

Dataset Description

The research is based on three key datasets: Reddit News, which compiles news headlines from Reddit; Combined News DJIA, a consolidated dataset of news headlines with an indicator on the price movement of DJIA; and DJIA Table, which provides historical performance data for the DJIA. The main focus will be on the Combined News DJIA and DJIA Table which have the necessary features and labels to conduct the research. A key column to highlight is the Label column in DJIA Table, which is a binary indicator: a value of 1 signifies that the DJIA price increased or remained the same, while a value of 0 indicates a decrease.

Data Preprocessing

To prepare the datasets for analysis, several preprocessing steps are applied. Firstly, data cleaning is performed by converting all news headlines to lowercase, removing stopwords, and eliminating punctuations and special characters to standardize the text. Tokenization is then performed to split the news headlines into individual tokens. Lemmatization is applied using NLTK's WordNetLemmatizer, which reduces words to their base forms while leaving those unchanged that are not found in the WordNet corpus. Once the text preprocessing is complete, the news headline dataset is merged with the stock price dataset, enabling the integration of textual features with a corresponding stock market movement for the modeling process. The combined dataset is then split into two subsets: 80% for training and 20% for testing, ensuring reliable model evaluation.

Methods Description

Feature extraction and engineering techniques were applied to prepare the text data for analysis. TF-IDF (Term Frequency-Inverse Document Frequency) was used to transform textual data into numerical features by emphasizing term importance, while sentiment scores—Positive, Negative, Neutral, and Compound—were generated using the VADER model to capture the emotional tone of the text. Named Entity Recognition (NER) was employed to extract meaningful entities, such as organizations, adding further context to the data. Lagged sentiment features were introduced to identify temporal patterns and improve predictive accuracy. Machine learning models, including Random Forest with SMOTE to address class imbalance, Logistic Regression with hyperparameter tuning, Naive Bayes incorporating NER features, and Linear Regression to infer relationships between features and DJIA, were implemented. To address high dimensionality, Principal Component Analysis (PCA) was applied to streamline the feature space while retaining essential information for model training and evaluation.

Model Development

Task 1: Classifications with TF-IDF and Sentiment

1. Initial Modeling: Random Forest

The modeling process began with the Random Forest classifier. Random Forest is an ensemble method that constructs multiple decision trees using random subsets of the data. The final prediction aggregates the outputs of all trees, helping reduce overfitting compared to a single decision tree. For the initial model, TF-IDF features (derived from news headlines) and sentiment scores (positive, negative, neutral, and compound) were used as input. However, the results indicated that the model struggled to predict stock decreases (Label 0):

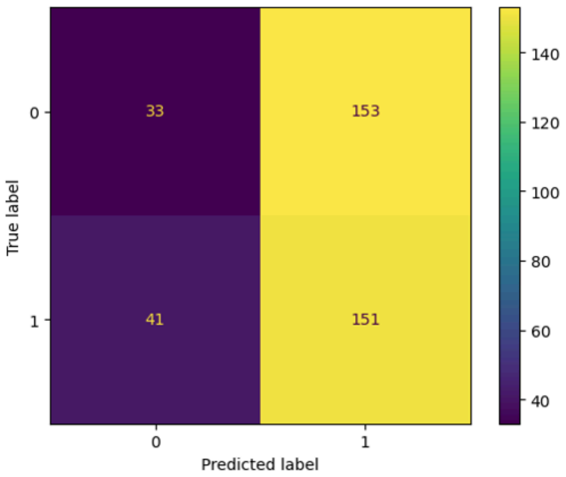


Figure 1: Confusion Matrix

- The Confusion Matrix revealed that predictions were heavily biased toward Label 1 (stock increases).
- **Performance Metrics**:
  - Accuracy: 0.49 (indicating poor overall performance).
  - Recall for Label 0: 18% (very low, meaning the model rarely identified decreases correctly).
  - ROC-AUC Score: 0.48 (worse than random guessing, which would yield 0.50).

These results pointed to two key issues:

1. Data Imbalance: Fewer samples for Label 0 led to a prediction bias.
2. Model Limitations: The Random Forest struggled to distinguish between increases and decreases effectively.

2. Model Improvement Strategies

2.1 Addressing Data Imbalance with SMOTE

To overcome the imbalance between the two classes, we applied SMOTE (Synthetic Minority Oversampling Technique). SMOTE generates synthetic samples for the minority class (Label 0) instead of duplicating existing ones, thereby balancing the dataset.

Outcome:
The model became more balanced in treating both classes, and there was a slight improvement in recall for Label 0.

2.2 Adding Lagged Features for Temporal Context

Given that financial markets often respond to previous trends, we introduced lagged features to the dataset. These included the sentiment scores (positive, negative, neutral, and compound) from the previous day to provide the model with historical context.

Outcome:
The addition of lagged features helped the model capture temporal relationships, slightly improving performance in predicting stock movements.

2.3 Hyperparameter Tuning with GridSearch

To optimize the Random Forest model, we used GridSearchCV, which systematically tests multiple combinations of parameters (e.g., tree depth, number of trees, and class weights). By evaluating performance across a range of settings, GridSearch identifies the best parameters for the model.

Best Parameters Identified:

- Number of estimators: 200
- Tree depth: 20
- Class weight: Balanced

Outcome:
GridSearch improved the balance between precision and recall, particularly for Label 0.

3. Logistic Regression: L1 vs L2 Regularization

To address the high dimensionality of TF-IDF features, we applied Logistic Regression with two types of regularization:

1. **L1 Regularization**: This shrinks some feature coefficients to zero, effectively performing feature selection and removing less important predictors.
2. **L2 Regularization:** This reduces all feature coefficients but does not eliminate any (no shrinking to zero). It works well when all features contribute meaningfully to predictions.

Model Comparison Results:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression (L1)	0.50	0.50	0.50	0.49	0.48
Logistic Regression (L2)	0.47	0.47	0.47	0.47	0.44

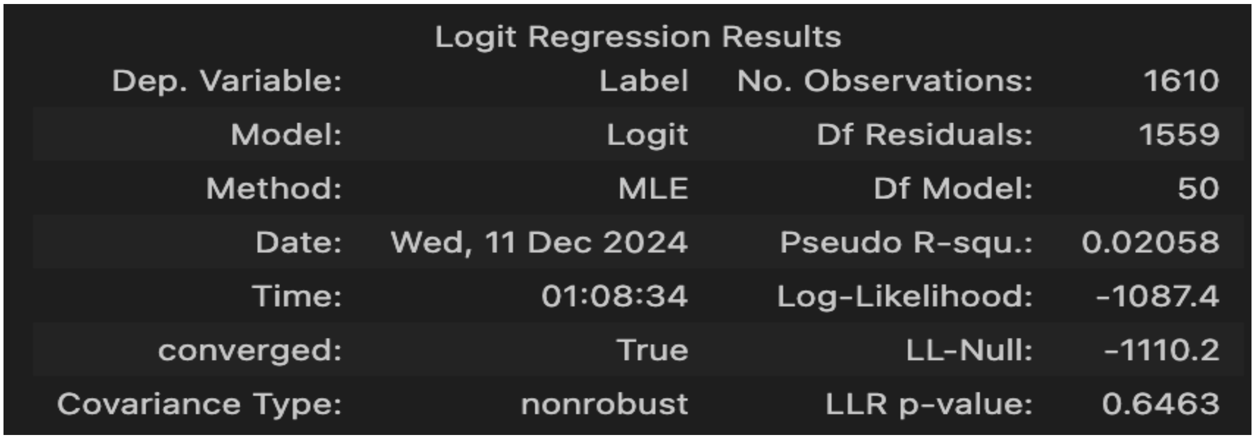
Observations:

- **L1 Regularization** outperformed L2 by removing irrelevant features and reducing the noise caused by high-dimensional TF-IDF data.
- L2 Regularization retained all features, which led to slightly lower performance due to overfitting on less relevant predictors.

The feature importance plot from the L1 model highlighted the most influential terms (positive and negative) for predicting stock price movements.

4. Principal Component Analysis (PCA)

To further simplify the high-dimensional TF-IDF features, we applied Principal Component Analysis (PCA). PCA reduced dimensionality by transforming features into new components that retain the maximum variance in the data.



```
                        Logit Regression Results
   Dep. Variable:              Label   No. Observations:       1610
          Model:               Logit       Df Residuals:       1559
         Method:                 MLE           Df Model:         50
           Date:    Wed, 11 Dec 2024     Pseudo R-squ.:     0.02058
           Time:            01:08:34     Log-Likelihood:    -1087.4
      converged:                True            LL-Null:    -1110.2
Covariance Type:           nonrobust        LLR p-value:     0.6463
```

Figure 2: Logistic Regression Results

Results:

- PCA successfully reduced the number of features, but only a few principal components were statistically significant.
- Despite dimensionality reduction, the model's performance did not improve:
  - R-squared values were low as well as high p-value.
  - The model struggled to identify patterns in the principal components.

Conclusion: The limited success of PCA suggests that stock price movements are influenced by highly complex factors beyond those captured by news sentiment or TF-IDF features.

5. Final Evaluation

After applying SMOTE, lagged features, GridSearch, logistic regression with L1/L2 regularization, and PCA, we compared the results:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest (Initial)	0.49	0.47	0.48	0.43	0.48
Random Forest (Improved with SMOTE)	0.49	0.51	0.51	0.49	0.48
Logistic Regression (L1)	0.50	0.50	0.50	0.49	0.48

Key Findings:

- **Random Forest** improved slightly after applying SMOTE and lagged features but still exhibited limited predictive power.
- **Logistic Regression (L1)** performed better by efficiently handling the high-dimensional
- **PCA** did not lead to significant improvements, indicating that the complexity of stock price movements cannot be captured by dimensionality reduction alone.

Task 2: Naive Bayes with Named Entities

Engelberg and Parsons (2011) found that certain named entities in the news were correlated with stock price movements. They demonstrated that mentions of influential entities, such as companies, executives, or key political figures, in news articles could significantly impact market participants' decisions, thereby affecting stock prices. Therefore, in the third model, Named Entities were extracted from news headlines, focusing on three key categories: Organization, GPE (Geopolitical Entity), and Person. The frequency distribution revealed that the top 100 entities in each category accounted for the majority of mentions. High-frequency entities such as Obama, China, Israel, and Putin highlighted their prominence in global discussions, underscoring their potential influence on market sentiment. The objective of this model is to discover whether the appearance of these entities in news articles could signal stock price movements.

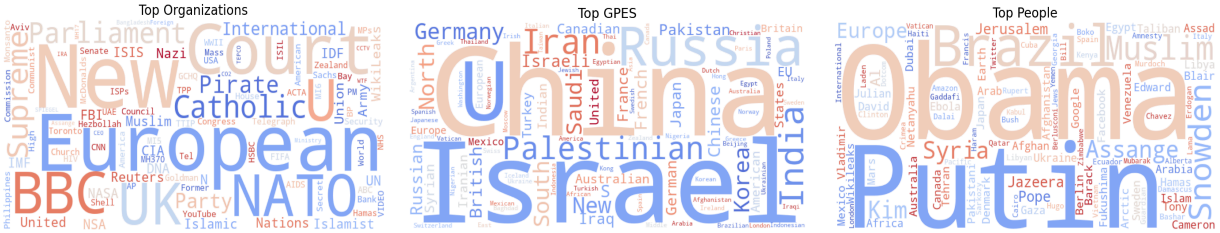


Figure 3: Top Named Entities of Organization, GPES, and People

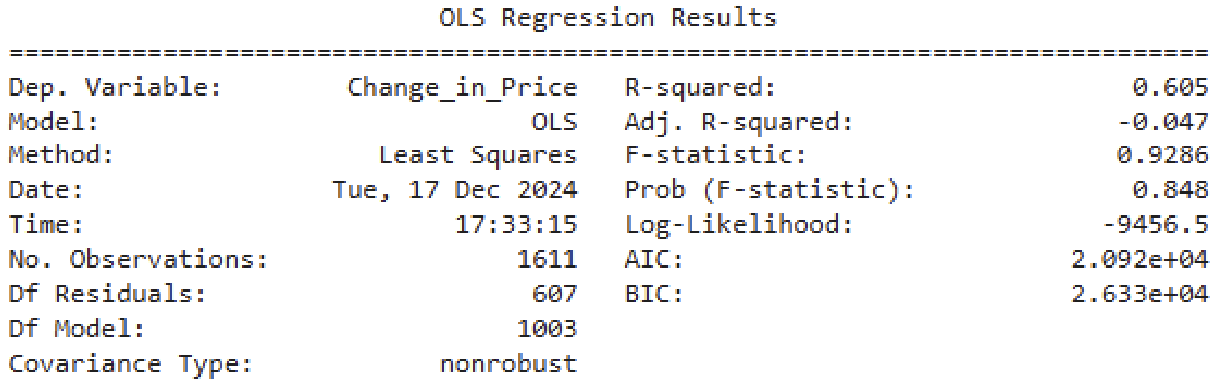
To explore this, a Naïve Bayes model was trained using whether news headlines contained the top 100 entities from each category as features. The results indicated that headlines mentioning entities such as the Foreign Ministry or the Security Council were more frequently associated with DJIA increases, while mentions of entities like ACTA and the European Court were more often linked to DJIA declines. These findings suggest that specific entities in the news may serve as early signals of market sentiment. However, despite discovering several named entities associated with stock price movements, the model's performance was not optimal, with an F1 score of 0.55.

One possible explanation is that news involving organizational entities, such as NASA or YouTube, tends to have a direct impact on the stock prices of the respective organizations or industries, but it is less likely to affect the entire DJIA. This is because the stock market is influenced by a wide range of factors, and the effect of individual entities may be more localized. Entities with broader political or geopolitical significance—such as governments or major international organizations—are more likely to impact market sentiment, which can trigger more significant shifts in the overall market index. This could explain why news about many entities fails to produce substantial movements in the DJIA.

Task 3: Linear Regression with TF-IDF and Sentiment

In addition to classification models, linear regression is also implemented to explore any potential inferences from the TF-IDF and the sentiment scores given that the classification models have poor performance in testing. The features used for the Linear Regression are the normalized TF-IDF, sentiment scores, and lagged sentiment scores. The normalized TF-IDF scores will range from 0 to 1. Using the statsmodel library, a linear regression model was generated with the R^2 value of 0.605 meaning 60.5%

of the variance in Change in Price can be explained by the features in the model. Looking beyond the R^2 value, however, it can be concluded that the R^2 value might be inflated given that the adjusted R^2 value is a negative value. The negative value indicates that the amount of features included in the linear regression model is large relative to the sample size with the 1003 features and only 1611 entries. In addition, a large difference between the Df Model of 1003 and Df Residuals (607) is a sign of overfitting. To overcome this issue, a form of dimensional reduction is needed.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Change_in_Price   R-squared:                       0.605
Model:                               OLS   Adj. R-squared:                 -0.047
Method:                    Least Squares   F-statistic:                    0.9286
Date:                   Tue, 17 Dec 2024   Prob (F-statistic):              0.848
Time:                           17:33:15   Log-Likelihood:                -9456.5
No. Observations:                   1611   AIC:                         2.092e+04
Df Residuals:                        607   BIC:                         2.633e+04
Df Model:                           1003
Covariance Type:               nonrobust
```

Figure 4: Linear Regression Result

Before implementing any model improvement measures, the top five words with the highest positive coefficients and the top 5 words with negative coefficients with p-values less than 0.05 are identified to explore potential meaningful insights from the model. This indicates that in this version of the linear model, these 10 words are statistically significant in predicting the Change in Price. One of the notable significant words was "Hacking" which had a coefficient of -215.540207 with a p-value of 0.014834. This means that an increase of 0.1 in TF-IDF value for the "Hacking" would reflect a negative change of 21.55 dollars to the price of that particular day. In a visual context, a line plot of TF-IDF value for "Hacking" and a line plot of Change in Price overtime was created. One particular point in time that stands out is during mid-2011 when the word "Hacking" was of high importance in daily news headlines. The negative sentiment of the word "Hacking" was further reflected by the significant negative change in price for that time frame.

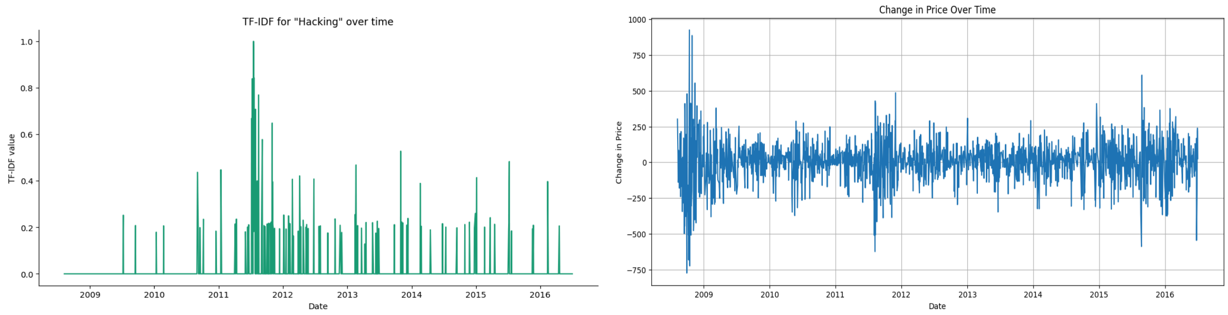


Figure 5: Example of "Hacking"

Future steps for improving the model include conducting Principal Component Analysis to reduce dimensionality, performing Correlation Analysis to identify and remove redundant features, and checking for Multicollinearity to address highly correlated predictors.

Model Evaluation

To evaluate the performance of the models, various metrics were considered, including accuracy, precision, recall, F1 score, and ROC-AUC. The classification models—Random Forest, Logistic Regression, and Naive Bayes—were assessed based on their ability to predict stock price movements. While the models showed some potential, their performance was limited, as evidenced by the moderate F1 scores. One key factor contributing to this limitation is the inherent complexity of stock price fluctuations, which are influenced by a wide range of factors beyond just news sentiment, including economic indicators, market trends, and investor behavior. While news provides valuable insights, it cannot fully capture the multitude of variables that drive stock market dynamics.

<table>
<tr><th colspan="6">Classification Models (for Prediction)</th></tr>
<tr><th>Models</th><th>Accuracy</th><th>Precision</th><th>Recall</th><th>F1-score</th><th>ROC-AUC</th></tr>
<tr><td>Random Forest</td><td>0.49</td><td>0.52</td><td>0.71</td><td>0.60</td><td>0.48</td></tr>
<tr><td>Logistic Regression</td><td>0.47</td><td>0.48</td><td>0.52</td><td>0.50</td><td>0.44</td></tr>
<tr><td>Naive Bayes</td><td>0.55</td><td>0.55</td><td>0.55</td><td>0.55</td><td></td></tr>
<tr><th colspan="6">Regression Model (for Inference)</th></tr>
<tr><th>Model</th><th>R2</th><th>Adj R2</th><th>F Statistic</th><th>Df Residuals</th><th>Df Models</th></tr>
<tr><td>Linear Regression</td><td>0.61</td><td>-0.047</td><td>0.93</td><td>607</td><td>1003</td></tr>
</table>

Figure 6: Model Evaluation Table

In addition to the classification models, the linear regression model provided valuable insights into how sentiment and specific keywords influence DJIA. By analyzing the coefficients of the regression model, it was observed that certain words related to geopolitical issues, natural disasters, and health crises had a stronger influence on the DJIA. For instance, words like "cancer" and "hacking" were found to have a significant negative impact, indicating that such topics may trigger negative market sentiment.

The complexity of market dynamics, combined with the presence of numerous external factors, suggests that while news sentiment can provide some indication of market trends, it is not sufficient to make reliable predictions on its own. Therefore, further research and the inclusion of additional features, such as economic indicators and investor sentiment, are necessary to improve the predictive power of these models.

Practical Implication

This project holds significant implications across various areas. It enhances risk management by providing a deeper understanding of how specific words of news impact stock prices, allowing investors to adapt their strategies in real time and mitigate potential risks. It also contributes to algorithmic trading, where sentiment analysis can be integrated into trading algorithms to refine decision-making processes and optimize trade execution based on shifting market sentiment trends. Furthermore, financial analysts can leverage these insights to fine-tune their market predictions. This, in turn, leads to more informed and data-driven recommendations for clients, empowering them to make better investment decisions.

Conclusion

This study demonstrates that combining sentiment and textual features offers valuable insights into stock price movements, but it also faces significant limitations due to the inherent complexity of market behavior. Text alone is unable to fully capture the multitude of factors influencing stock trends, underscoring the need for more sophisticated approaches. To address these challenges, future work should explore the application of deep learning techniques, which can better model complex relationships, and integrate additional data sources, such as financial indicators, historical market data, and macroeconomic factors. This will help create more robust models, leading to improved predictions and more accurate decision-making in the financial sector.

Reference

Engelberg, J., & Parsons, C. A. (2011). The role of news in market participants' decision making. The Journal of Financial Economics, 99(3), 451-469.

Fama, E. F. (1997). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics, 49*(3), 283-306.

Data Source: https://www.kaggle.com/datasets/aaron7sun/stocknews