Stock Market News Sentiment Analysis & Summarization

Project #6- Introduction to Natural Language Processing

Saturday, January 4, 2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Executive Summary

This project aimed to develop an AI-driven sentiment analysis model to process stock-related news and assess its impact on market sentiment, supporting informed investment decisions. After evaluating multiple models, the **Tuned GloVe Model** was selected as the final solution due to its balanced performance across accuracy, precision, recall, and F1 score. Despite moderate generalization challenges observed during testing, the model demonstrated robust capability in handling the dataset's high dimensionality, imbalanced sentiment distribution, and relatively small sample size. Its simplicity, reliability, and shortest computation time make it well-suited for analyzing sentiment trends, offering actionable insights to financial analysts. Additionally, tools such as llama-cpp-python and huggingface_hub were successfully installed and integrated to support article summarization. Testing this functionality with the stock_news.csv database and making the necessary format adjustments produced the expected tabular outputs, further demonstrating the project's ability to automate sentiment extraction and provide meaningful summaries.

While the Tuned GloVe Model demonstrates strong performance for sentiment analysis, there are several areas for improvement to enhance its generalization and effectiveness. One key issue is the drop in performance on the test set, indicating challenges in generalization likely linked to overfitting, as implied by the perfect training scores across all models evaluated. This can be addressed by collecting additional labeled data, applying data augmentation techniques, and implementing advanced regularization methods. The class imbalance in the dataset, with neutral sentiment dominating, may bias the model's predictions; balancing the dataset through oversampling, undersampling, or class weighting during training could mitigate this issue. The dataset's high dimensionality, with 384 features, may also contribute to inefficiencies and overfitting; dimensionality reduction techniques such as feature selection can help retain the most relevant features. Exploring alternative embeddings like BERT could further improve contextual understanding, while ensemble models might offer greater robustness. Finally, expanding the test set and fine-tuning hyperparameters through grid search would ensure more robust evaluation of the model's effectiveness. Addressing these areas would improve generalization, sentiment classification accuracy, and actionable insights for stock price prediction and investment strategies.

Executive Summary- Observations

After evaluating the performances of each of the six models designed to summarize and classify news articles into their respective sentiment categories or positive, neutral, or negative, the following observations and actionable insights can be made as follows:

Root Problems with Data & Models

Overfitting:

- The model might have fit too closely to the training data. Indicators include perfect or near-perfect training scores but low test scores.

Data Issues:

- The training, validation, and test datasets might have different distributions.
  - Relatively small sample dataset overall
  - Imbalanced classes
  - Different language styles or sentiment signals in the test data.

Model Limitations:

- Gradient Boosting models can struggle with very high-dimensional data or subtle relationships present in text embeddings.

Executive Summary- Actionable Insights

Options to Improve Performance

1. Generalization to Test Data

- Issue: The model's performance drops on the test set compared to the validation set, indicating challenges in generalization.
- Improvement:
    - Collect additional labeled data to improve the model's exposure to diverse examples.
    - Use data augmentation techniques to create variations of existing samples.
    - Explore advanced regularization techniques to reduce overfitting.

2. Addressing Class Imbalance

- Issue: The dataset is imbalanced, with neutral sentiment dominating, which may bias the model's predictions.
- Improvement:
    - Apply oversampling for minority classes (positive and negative) or undersampling for the majority class (neutral).
    - Use class weighting during model training to balance the impact of each class on the loss function.

3. Dimensionality Reduction

- Issue: The dataset has 384 features, which may lead to overfitting and inefficiencies, especially given the small sample size.
- Improvement:
    - Perform feature selection to identify and keep only the most relevant features.

4. Experimenting with Alternative Models

- Issue: The GloVe embeddings, while effective, might not fully capture contextual nuances.
- Improvement:
    - Explore contextual embeddings like BERT which provides better handling of nuanced text.
    - Test ensemble models that combine predictions from multiple embedding techniques to improve robustness.

Executive Summary- Actionable Insights

Options to Improve Performance

5. Fine-Tuning Hyperparameters

- Issue: The current hyperparameters, while effective, might not fully optimize the model.
- Improvement:
    - Conduct a more exhaustive hyperparameter search using grid search to fine-tune parameters like learning rate, regularization strength, and feature selection thresholds.
    -

6. Robust Evaluation

- Issue: Small validation and test sets may not provide stable performance metrics.
- Improvement:
    - Expand the test set to ensure robust evaluation of the model.

Executive Summary- Recommendation

Recommendation

In conclusion, the **Tuned GloVe Model with Gradient Boost Classifier** demonstrated the strongest performance during validation, making it the most suitable model for analyzing stock-related news sentiment. Its balanced performance across accuracy, precision, recall, and F1 score aligns well with the goal of providing actionable insights for trading decisions. However, the drop in performance on the test set highlights the need for further improvements to enhance generalizability and robustness. By addressing overfitting, refining the dataset, and implementing regular monitoring and updates post-deployment, this model can effectively support more informed and accurate investment strategies. With these enhancements, the company will gain a reliable tool to better interpret market sentiment and drive competitive decision-making.

Business Problem Overview

&

Solution Approach

Business Problem Overview

In the fast-paced and highly competitive financial industry, stock prices are significantly influenced by factors such as financial performance, innovations, collaborations, and market sentiment. News and media reports play a pivotal role in shaping investor perceptions, often leading to rapid market fluctuations. However, the overwhelming volume of news and opinions from diverse sources makes it challenging for investors and financial analysts to stay informed and accurately interpret their impact on stock prices.

To address this challenge, an investment startup seeks to leverage artificial intelligence to analyze stock-related news and extract actionable insights. By developing an AI-driven sentiment analysis system, the goal is to process and summarize news articles at a weekly level, enabling more accurate stock price predictions and optimized investment strategies. This solution will empower financial analysts to make more informed decisions, providing a competitive edge in the dynamic stock trading environment.

Solution Approach

To address the challenge of analyzing stock-related news sentiment, an AI-driven solution was developed using advanced natural language processing techniques. The approach began with the preprocessing of historical news data, including cleaning, tokenization, and embedding generation using models such as Word2Vec, GloVe, and Sentence Transformers. These embeddings were paired with machine learning classifiers, including the classifier of choice, Gradient Boosting, to predict sentiment (positive, neutral, or negative) associated with stock performance. Hyperparameter tuning was conducted to optimize model performance, focusing on key metrics such as accuracy, precision, recall, and F1 score, which was determined to be the most appropriate metric to focus on to evaluate the model. The final model, a tuned Sentence Transformer with Gradient Boost Classifier, was selected based on its superior validation performance. Weekly sentiment summaries were generated to provide actionable insights, integrating these outputs with stock price trends to support investment strategy optimization. This systematic approach ensured the solution captured nuanced sentiment signals, aligning with the business objective of enhancing decision-making in the financial domain.

Exploratory Data Analysis

EDA Results- Univariate analysis

Sentiment label (*SEE UA1*)

- The dataset's sentiment distribution shows an imbalance, with 48% of articles classified as neutral, 25% as positive, and 27% as negative, reflecting the predominance of neutral or objective news in stock-related reporting. This imbalance may lead the model to over-predict neutral sentiment, necessitating strategies like class weighting or resampling to ensure balanced performance across all classes. Metrics such as precision, recall, and F1 score, particularly for the minority classes, should be prioritized over accuracy to evaluate the model effectively. Misclassifications between neutral and other sentiments could significantly impact stock price predictions, highlighting the importance of thorough error analysis. Additionally, incorporating features like article length or polarity scores may enhance the model's ability to distinguish between neutral and impactful sentiments. A robust classification of positive and negative sentiments is critical for actionable insights, ensuring the project supports accurate and informed stock market decisions.

Density Plot of Price [Open,High,Low,Close] (*SEE UA2*)

- The density plot of stock prices reveals a bimodal distribution, characterized by two distinct peaks. The main peak, steep and concentrated, occurs around $42-45, representing the majority of stocks. A smaller, flatter secondary peak appears near $65-67, suggesting a secondary group of stocks in a higher price range. Between these peaks lies a valley around $55-60, highlighting a price gap with fewer stocks trading in this range. The bulk of prices fall between $35-55, with the highest density observed in the $40-50 range, and only a few stocks trading below $35 or above $70. All four price metrics—High, Low, Open, and Close—closely track each other, with overlapping density curves indicating consistent relationships across metrics. Close prices exhibit slightly higher density at lower values, while the High and Low prices dominate the primary and secondary peaks. The smooth tails of the distribution reflect a natural tapering off at extreme price ranges. The secondary cluster at higher prices and the price gap around $55-60 warrant further investigation to understand the underlying factors shaping this structure.

EDA Results- Univariate analysis

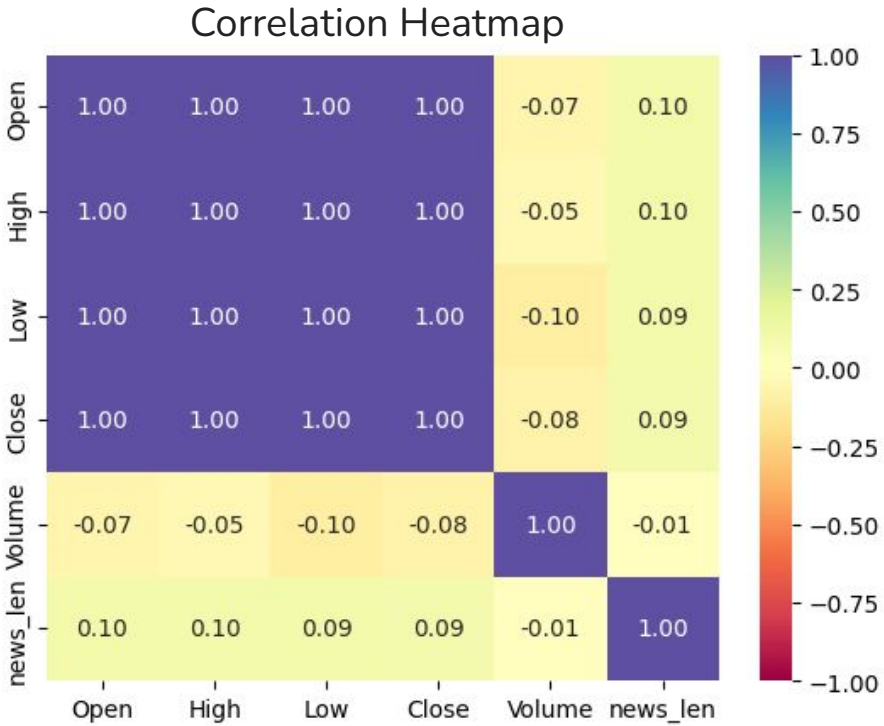
Volume *(SEE UA3)*

- The volume distribution is right-skewed, with most trading volumes concentrated between 75-150 million shares and a peak around 100-125 million shares. A long tail extends to higher volumes, with fewer stocks exceeding 200 million shares and very few below 50 million shares. The distribution shows clusters, including a main concentration at 100-125 million shares and a smaller secondary bump around 225-250 million shares. The median volume is approximately 115-120 million shares, highlighting significant variation in trading activity. The right skew indicates that a subset of stocks attracts significantly higher trading volumes, reflecting potential disparities in market interest and participation. These patterns suggest that trading volume is highly variable, with certain stocks driving much of the observed activity.

News Length *(SEE UA4)*

- The news length distribution is roughly normal with a slight negative skew, peaking around 50-55 words and concentrating between 45-55 words. Most articles range from 40-60 words, with very few exceeding this range or falling below 30 words. The distribution shows a sharp drop-off in frequency outside this range, with isolated cases of very short articles (20-30 words). The narrow range and consistency suggest that the content is likely already standardized, such as news summaries or headlines, rather than full articles. This uniformity indicates a focus on concise reporting, providing a clear and predictable dataset for analysis.

EDA Results- Bivariate Analysis



The heatmap analysis reveals strong positive correlations among price metrics (Open, High, Low, Close), indicating they move in perfect unison. Volume shows slight negative correlations with all price metrics, with the strongest being with Low prices (-0.10), suggesting higher volumes are associated with slightly lower prices. News length exhibits very weak positive correlations with price metrics (~0.09-0.10) and almost no correlation with volume (-0.01), indicating that news content length has minimal impact on price or trading activity. Overall, price metrics form a highly correlated cluster, while volume and news length appear to be largely independent variables. These patterns suggest that stock prices move in sync, while trading volume operates relatively independently of both price and news length.

EDA Results- Bivariate analysis

Great Learning
POWER AHEAD

Sentiment label vs Price *(SEE BA1)*

- The relationship between stock prices and sentiment shows systematic but modest patterns. Across all sentiment categories, the price metrics (Open, High, Low, Close) display similar trends, with the majority of prices concentrated in the $40-50 range and outliers in the $65-67 range. Positive sentiment (1) is associated with slightly higher median prices, while negative sentiment (-1) corresponds to slightly lower medians. Neutral sentiment (0) consistently falls between the two, suggesting a subtle gradient in price levels influenced by sentiment. Price variation appears consistent across sentiment labels, with similar interquartile ranges and symmetric boxplots, indicating relatively normal distributions. Close prices show slightly more variability compared to other metrics, and the presence of consistent outliers at higher price levels reflects occasional market volatility irrespective of sentiment. The medians exhibit a slight upward trend from negative to positive sentiment, reinforcing the weak but systematic relationship between sentiment and price. Overall, while positive sentiment is linked to marginally higher prices and negative sentiment to lower prices, the effect is not pronounced. This suggests that sentiment does play a role in influencing stock prices but is not the sole determinant, with other market factors likely driving more substantial price movements.

Sentiment label vs Volume *(SEE BA2)*

- The distribution of trading volumes across sentiment categories reveals notable patterns. Neutral sentiment (0) exhibits the highest median volume and the largest variability, with a broader interquartile range and higher upper quartile compared to negative (-1) and positive (1) sentiments. While all categories feature outliers in the 225-250 million share range, the median and interquartile ranges for negative and positive sentiments are similar, indicating comparable trading behaviors during these periods. Most trading volumes are concentrated between 75-175 million shares, with minimum volumes ranging from 50-75 million shares across all sentiments. Despite the variability, sentiment appears to have a limited impact on trading volume. Neutral news periods display greater volume variability, suggesting a higher degree of uncertainty or diverse market reactions. In contrast, trading volumes are more consistent during positive and negative news periods, implying steadier market participation. The presence of outliers across all sentiment categories further underscores the complex relationship between sentiment and trading volume, suggesting that while sentiment impacts stock prices, its influence on trading volume is less direct.

EDA Results- Bivariate analysis

Date vs Price *(SEE BA3)*

- The stock price during the observed period exhibited a range of $35 to $67, with several significant spikes and four major peaks in mid-January ($63), mid-February ($67), early March ($65), and early April ($65). These peaks were characterized by rapid price increases followed by sharp declines, highlighting a recurring pattern of volatility. Major price spikes occurred approximately once a month, with recovery periods averaging two weeks, reflecting a predictable temporal rhythm. Early 2019 showed greater volatility, with wider daily price ranges and increased gaps between high and low prices, particularly during periods of sharp rises and declines. High and low prices closely tracked open and close values, with close prices often lower during downward trends. Despite the volatility, the stock demonstrated a general upward trajectory from ~$40 in January to ~$50 in April, with the final period showing more stability at the higher price level. This pattern of sharp fluctuations followed by consolidation suggests potential stabilization and a shift toward reduced volatility in the later months.

Volume vs Close price *(SEE BA4)*

The relationship between trading volume and stock price reveals clear patterns of market behavior and trading dynamics. Major price spikes consistently align with volume spikes, with peak trading volumes typically occurring in the range of 200-250 million shares, while baseline volumes range between 75-150 million shares. High-volume events often signal price reversals, indicating significant market activity and potential institutional involvement. Notable price increases, such as those in mid-January, early February, and early March to early April, coincided with sharp volume surges, reinforcing the connection between momentum-driven trading and price movements. Volume spikes are typically short-lived (1-2 days) and often mark transitional phases in price trends. Conversely, lower trading volumes are associated with price stability, suggesting reduced market interest during consolidation periods. The final observed period (mid-April onward) reflects increased price stability, accompanied by moderate and consistent trading volume, signaling a potential shift toward equilibrium. Overall, volume serves as a critical indicator of market participation and can provide insights into future price direction.

Data Preprocessing

Data Preprocessing

Duplicate value check

- There were 0 duplicate values in the dataset.

Missing value check

- There were 0 missing values in the dataset.

Data preprocessing for modeling

- printed the statistical summary of the 'Date' column
- pick the 'Label' column as the target variable
- Train-validation-test split
    - X_train = select all rows where the 'Date' is before '2019-04-01'
    - X_val = select all rows where the 'Date' is from '2019-04-01 to '2019-04-16' (excluded)
    - X_test = select all rows where the 'Date' is from '2019-04-16' till the end

Model Performance Summary

Sentiment Analysis - Model Evaluation Criterion

```
"""
    Compute various performance metrics for a classification model using sklearn.

    Parameters:
    model (sklearn classifier): The classification model to evaluate.
    predictors (array-like): The independent variables used for predictions.
    target (array-like): The true labels for the dependent variable.

    Returns:
    pandas.DataFrame: A DataFrame containing the computed metrics (Accuracy, Recall, Precision, F1-score). """
```

Metrics scoring criteria= F1 score

Given the small sample size, class imbalance could greatly influence results. F1 Score is robust and accounts for both precision and recall, making it a good choice to balance the trade-offs between false positives and false negatives across classes.

Sentiment Analysis - Model Building

Overview of Word2Vec base model

- Length of the vocabulary is 4,682 words.
- Time taken to create a dataframe of the vectorized documents: 0.5076971054077148 seconds.
- X_train_wv.shape, X_val_wv.shape, X_test_wv.shape

  (286, 300) (21, 300) (42, 300)

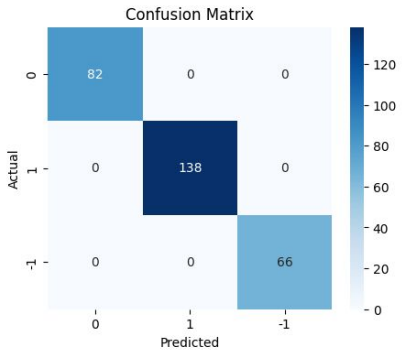
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Observations:

- Shortest computation time of base models
- Perfect training score
- Comparatively small vocabulary database
- Very high dimensionality of features can lead to overfitting, computational complexity

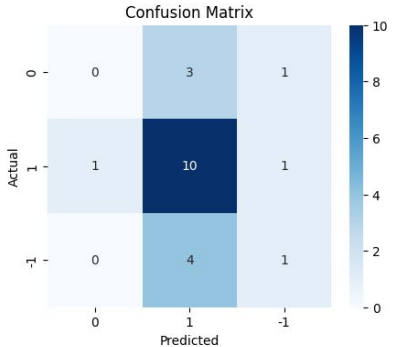
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.52381	0.52381	0.4155	0.453612

Sentiment Analysis - Model Building

Overview of GloVe base model

- Length of the vocabulary is 400,000 words.
- Time taken to create a dataframe of the vectorized documents: 35.96309995651245 seconds.
- X_train_gl.shape, X_val_gl.shape, X_test_gl.shape

  (286, 100) (21, 100) (42, 100)

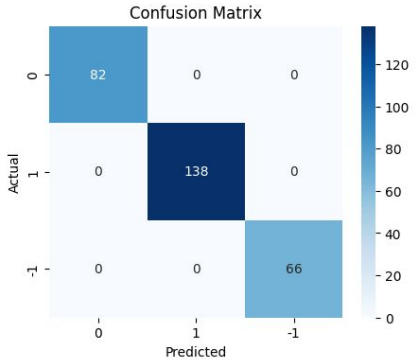
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Observations:

- Longest computation time of base models
- Perfect training score
- Larger of two vocabulary databases
- High dimensionality of features can lead to overfitting, computational complexity

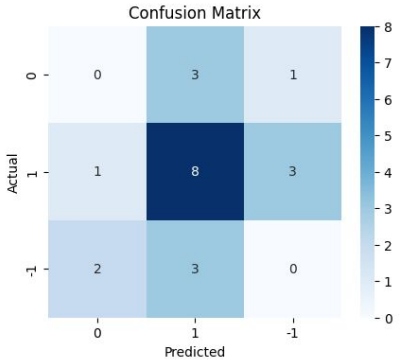
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.380952	0.380952	0.326531	0.351648

Sentiment Analysis - Model Building

Overview of Sentence Transformer base model

- Length of the vocabulary is 400,000 words.
- Time taken to encode the “News” column of the dataset: 2.5004961490631104 seconds.
- X_train_st.shape, X_val_st.shape, X_test_st.shape

  (286, 384) (21, 384) (42, 384)

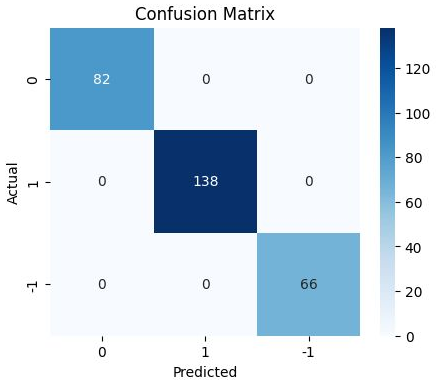
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Observations:

- Median computation time of base models
- Perfect training score
- Larger of two vocabulary databases
- Very high dimensionality of features can lead to overfitting, computational complexity

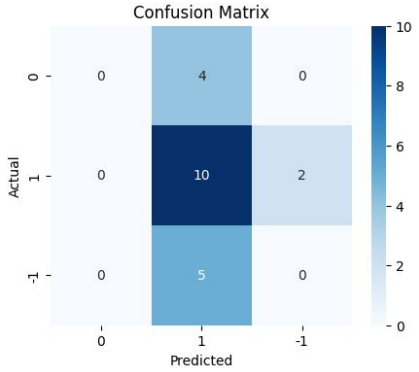
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.300752	0.368664

Sentiment Analysis - Model Improvement

Overview of Tuned Word2Vec model

- Time taken to create a dataframe of the vectorized documents: 619.8892052173615 seconds.
- GridSearchCV
- Parameters:
  - 'max_depth': np.arange(3,7),
  - 'min_samples_split': np.arange(5,12,2),
  - 'max_features': ['log2', 'sqrt', 0.2, 0.4]

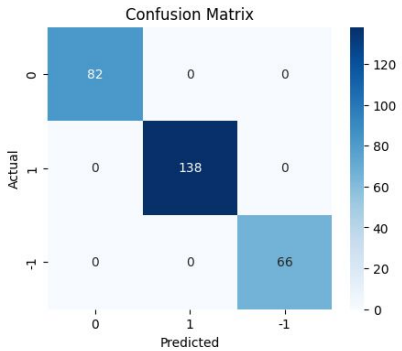
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(max_features='sqrt', min_samples_split=9,
                           random_state=42)
```

Observations:

- Second shortest computation time of tuned models
- Perfect training score
- Very high dimensionality of features can lead to overfitting, computational complexity

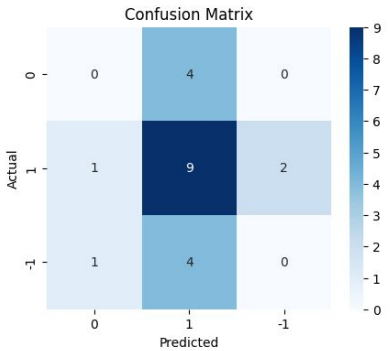
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.428571	0.428571	0.302521	0.35468

Sentiment Analysis - Model Improvement

Overview of Tuned GloVe model

- Time taken to create a dataframe of the vectorized documents: 313.8208005428314 seconds.
- GridSearchCV
- Parameters:
  - 'max_depth': np.arange(3,7),
  - 'min_samples_split': np.arange(5,12,2),
  - 'max_features': ['log2', 'sqrt', 0.2, 0.4]

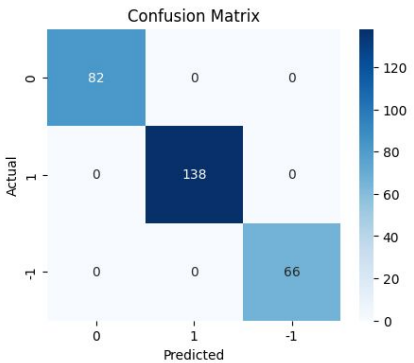
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(max_features='log2', min_samples_split=9,
                           random_state=42)
```

Observations:

- Shortest computation time of tuned models
- Perfect training score
- High dimensionality of features can lead to overfitting, computational complexity

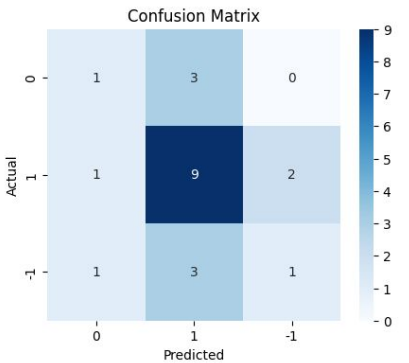
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.52381	0.52381	0.485714	0.494898

Sentiment Analysis - Model Improvement

Overview of Tuned Sentence Transformer model

- Time taken to create a dataframe of the vectorized documents: 768.1899855136871 seconds.
- GridSearchCV
- Parameters:
  - 'max_depth': np.arange(3,7),
  - 'min_samples_split': np.arange(5,12,2),
  - 'max_features': ['log2', 'sqrt', 0.2, 0.4]

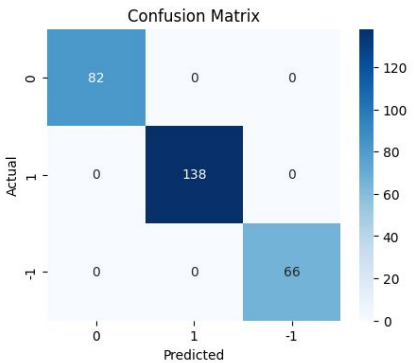
Chosen classifier

```
GradientBoostingClassifier
GradientBoostingClassifier(max_features='sqrt', min_samples_split=11,
                           random_state=42)
```

Observations:

- Longest computation time of tuned models
- Perfect training score
- High dimensionality of features can lead to overfitting, computational complexity

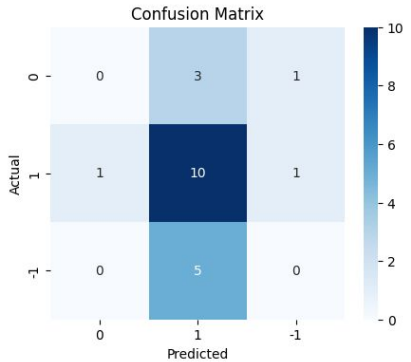
Training matrix



Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation matrix



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.31746	0.380952

Sentiment Analysis - Model Improvement

Training performance comparison:

	Base Model (Word2Vec)	Base Model (GloVe)	Base Model (Sentence Transformer)	Tuned Model (Word2Vec)	Tuned Model (GloVe)	Tuned Model (Sentence Transformer)
Accuracy	1.0	1.0	1.0	1.0	1.0	1.0
Recall	1.0	1.0	1.0	1.0	1.0	1.0
Precision	1.0	1.0	1.0	1.0	1.0	1.0
F1	1.0	1.0	1.0	1.0	1.0	1.0

Word2Vec

BASE

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.52381	0.52381	0.4155	0.453612

TUNED

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.428571	0.428571	0.302521	0.35468

- Accuracy: Decreased .10
- Recall: Decreased .10
- Precision: Decreased .11
- F1: Decreased .10

Model performance decrease in 100% of metrics

GloVe

BASE

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.380952	0.380952	0.326531	0.351648

TUNED

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.52381	0.52381	0.485714	0.494898

- Accuracy: Increased .14
- Recall: Increased .14
- Precision: Increased .16
- F1: Increased .14

Model performance increase in 100% of metrics

Sentence Transformer

BASE

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.300752	0.368664

TUNED

Validation performance:

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.31746	0.380952

- Accuracy: No change
- Recall: No change
- Precision: Increased .01
- F1: Increased .02

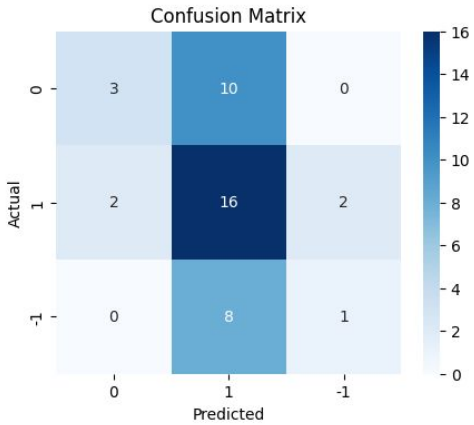
Model performance increase in 50% / decrease in 50% of metrics

Sentiment Analysis – Model Performance Comparison

Validation performance comparison:

	Base Model (Word2Vec)	Base Model (GloVe)	Base Model (Sentence Transformer)	Tuned Model (Word2Vec)	Tuned Model (GloVe)	Tuned Model (Sentence Transformer)
Accuracy	0.523810	0.380952	0.476190	0.428571	0.523810	0.476190
Recall	0.523810	0.380952	0.476190	0.428571	0.523810	0.476190
Precision	0.415500	0.326531	0.300752	0.302521	0.485714	0.317460
F1	0.453612	0.351648	0.368664	0.354680	0.494898	0.380952

Training matrix



Test performance for the final model:

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.481232	0.421076

Sentiment Analysis – Final Model

Given the dataset's high dimensionality, imbalance, and relatively small sample sizes, the **Tuned GloVe Model** stands out as the most robust and reliable option for generalizing across sentiments. With the shortest computation time of the three tuned models, it achieves strong overall performance with the highest F1 score and a balanced accuracy and recall, all without excessive complexity, making it well-suited to handle the dataset constraints. This performance demonstrates its suitability for delivering actionable insights into stock-related news sentiment, aligning effectively with the project's business objectives.

The performance of the tuned GloVe model on the test data demonstrates moderate generalization, with an accuracy of 47.6%, recall of 47.6%, precision of 48.1%, and an F1 score of 42.1%. While the model shows balanced performance between precision and recall, the drop in F1 score from the validation phase (49.5%) indicates challenges in fully generalizing to unseen data. This suggests the model may have slightly overfit the training data, limiting its ability to consistently capture sentiment signals in new examples. The close alignment between precision and recall highlights the model’s attempt to balance false positives and false negatives, but the overall decline in performance underscores the need for further improvements. Addressing class imbalance, augmenting the dataset, or refining hyperparameters could enhance its ability to generalize effectively. Despite these limitations, the model provides a reasonable foundation for sentiment classification in the context of this project.

Content Summarization – Data Preprocessing

Installing and importing the necessary libraries

- llama-cpp-python
- huggingface_hub

Loading the data

- stock_news.csv

Content Summarization – Modeling Approach

Loading the LLM Model (Llama-ccp-python)

model_name_or_path = "TheBloke/Mistral-7B-Instruct-v0.2-GGUF"

model_basename = "mistral-7b-instruct-v0.2.Q6_K.gguf"

```
model_path = hf_hub_download(
repo_id='TheBloke/Mistral-7B-Instruct-v0.2-GGUF'
filename="mistral-7b-instruct-v0.2.Q6_K.gguf"

llm = Llama(
model_path=model_path,
n_gpu_layers=100,
n_ctx=4500,
)
```

AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 | BLAS = 1 | SSE3 = 1 | SSSE3 = 1 | VSX = 0 |

Content Summarization – Sample Input

1. **Role**: Specifies the role the LLM will be taking up to perform the specified task, along with any specific details regarding the role
2. **Task**: Specifies the task to be performed and outlines what needs to be accomplished, clearly defining the objective
3. **Instructions**: Provides detailed guidelines on how to perform the task, which includes steps, rules, and criteria to ensure the task is executed correctly
4. **Output Format**: Specifies the format in which the final response should be structured, ensuring consistency and clarity in the generated output

You are an expert data analyst specializing in news content analysis.

Task: Analyze the provided news headline and return the main topics contained within it.

Instructions:

1. Read the news headline carefully.

2. Identify the main subjects or entities mentioned in the headline.

3. Determine the key events or actions described in the headline.

4. Extract relevant keywords that represent the topics.

5. List the topics in a concise manner.

Return the output in JSON format with keys as the topic number and values as the actual topic.

Content Summarization – Sample Output

- As an output, the model is expected to return a JSON containing two keys, one for Positive Events and one for Negative Events.
- Return the output in JSON format with keys as the topic number and values as the actual topic.
  - Example: {"1": "Politics", "2": "Economy", "3": "Health" }

Defining the response function:

```
model_output = llm(
F"""
[INST]
{prompt}
News Articles: {news}
[/INST]
 """,
max_tokens=256,
temperature=0.7,
top_p=0.9,
top_k=50,
echo=False,
)
final_output = model_output["choices"][0]["text"]
return final_output
```

Content Summarization – Raw Model Output

```
        {
          "topic_1": "Apple's Q1 revenue warning and its impact on tech stocks",
          "keywords": ["Apple", "Q1 revenue warning", "tech stocks", "notable suppliers"]
        },
        {
          "topic_2": "Global economic concerns following Apple's revenue warning",
          "keywords": ["global economy", "weak economic data", "US-China trade tensions", "Apple's underperformance"]
        },
        {
          "topic_3": "Impact of Apple's Q1 warning on currency markets (Japanese yen and USD JPY)",
          "keywords": ["Apple", "Q1 revenue warning", "USD JPY", "Japanese yen"]
        },
        {
          "topic_4": "Impact of global economic concerns on long-term US Treasury securities",
          "keywords": ["long-term U.S. Treasury securities", "yields at lowest levels in nearly a year", "global economy"]
        },
        {
          "topic_5": "Roku
CPU times: user 21.7 s, sys: 5.25 s, total: 26.9 s
Wall time: 28.1 s
```

Content Summarization – Final Output

1. **Formatting the Model Output**

data_1['model_response_parsed'] = data_1['Key Events'].apply(extract_json_data)

	Date	News	Key Events	model_response_parsed
0	2019-01-06	The tech sector experienced a significant dec...	[\n {\n "topic_1": "Apple's Q...	{'topic_1': 'Apple's Q1 revenue warning and it...
1	2019-01-13	Sprint and Samsung plan to release 5G smartph...	1. 5G smartphones: Sprint, Samsung, nine U.S. ...	{}
2	2019-01-20	The U.S. stock market declined on Monday as c...	[\n 1. "U.S. stock market decline",\n ...	{}
3	2019-01-27	The Swiss National Bank (SNB) governor, Andre...	{\n "1": "Swiss National Bank (SNB) a...	{}
4	2019-02-03	Caterpillar Inc reported lower-than-expected ...	{\n "1": "Apple: Lower-than-expected...	{}

model_response_parsed = pd.json_normalize(data_1['model_response_parsed'])

	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	1	2	...	5	6	7	8	9	10	11	12	13	14
0	Apple's Q1 revenue warning and its impact on t...	Global economic concerns following weak econom...	Impact of Apple's Q1 revenue warning on curren...	Apple's underperformance in Q1 and its cause (...	Risk aversion and market declines following Ap...	US-China trade tensions as a factor in Apple's...	Increase in demand for safe haven assets (Japa...	Decelerating factory activity in China and the...	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2. **Concatenate the original data_1 and model_response_parsed horizontally**

final_output = pd.concat([data_1, model_response_parsed], axis=1)

3. **Print the shape and columns**

```
(18, 26)
Index(['Date', 'News', 'Key Events', 'model_response_parsed', 'topic_1',
       'topic_2', 'topic_3', 'topic_4', 'topic_5', 'topic_6', 'topic_7',
       'topic_8', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11',
       '12', '13', '14'],
      dtype='object')
```

Content Summarization – Final Output

4. **Replace with actual column names from the output of the print statement**

   final_output = data_1[['Date', 'News', 'Key Events', 'model_response_parsed']].copy()

   final_output.columns = ['Week End Date', 'News', 'Week Positive Events', 'Week Negative Events']

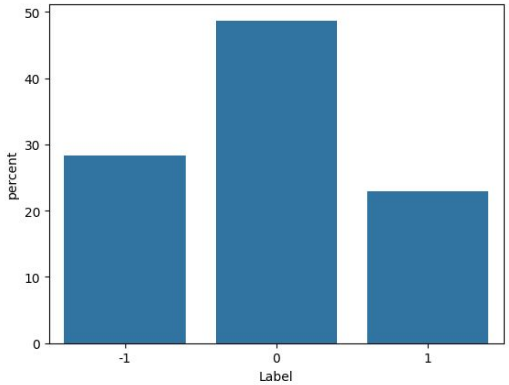
5. **Assign the modified column names back to final_output**

	Week End Date	News	Week Positive Events	Week Negative Events	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	...	5	6	7	8	9	10	11	12	13	14
0	2019-01-06	The tech sector experienced a significant dec...	Apple's Q1 revenue warning and its impact on t...	Global economic concerns following weak economi...	Impact of Apple's Q1 revenue warning on curren...	Apple's underperformance in Q1 and its cause (...	Risk aversion and market declines following Ap...	US-China trade tensions as a factor in Apple's...	Increase in demand for safe haven assets (Japa...	Decelerating factory activity in China and the...	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2019-01-13	Sprint and Samsung plan to release 5G smartph...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2019-01-20	The U.S. stock market declined on Monday as c...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2019-01-27	The Swiss National Bank (SNB) governor, Andre...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2019-02-03	Caterpillar Inc reported lower-than-expected ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

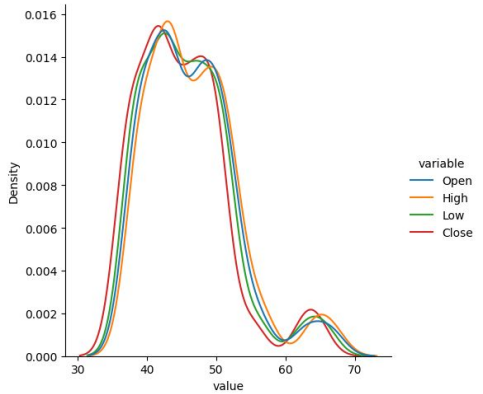
APPENDIX

EDA Results- Univariate Analysis

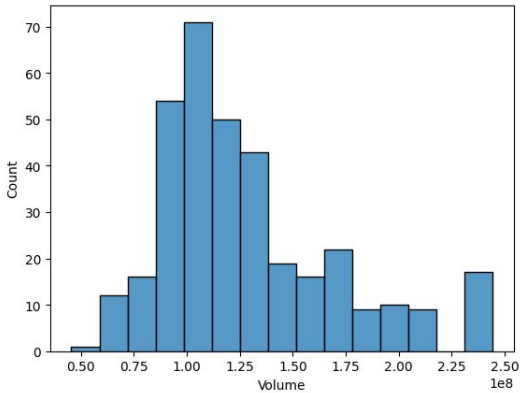
UA1- Sentiment label



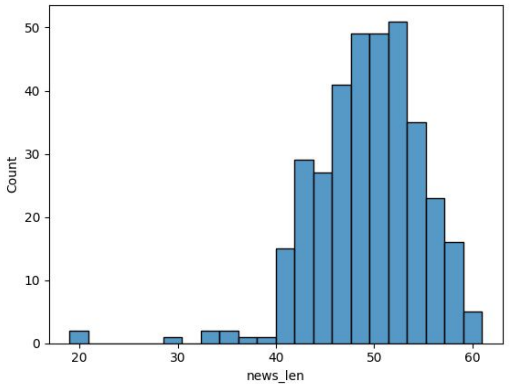
UA2- Density plot of price



UA3- Volume

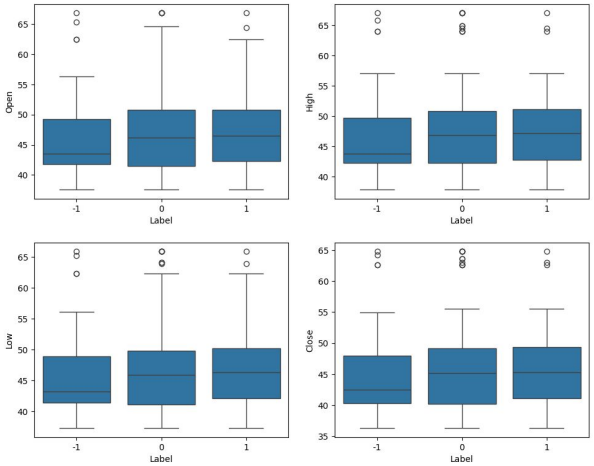


UA4- News length

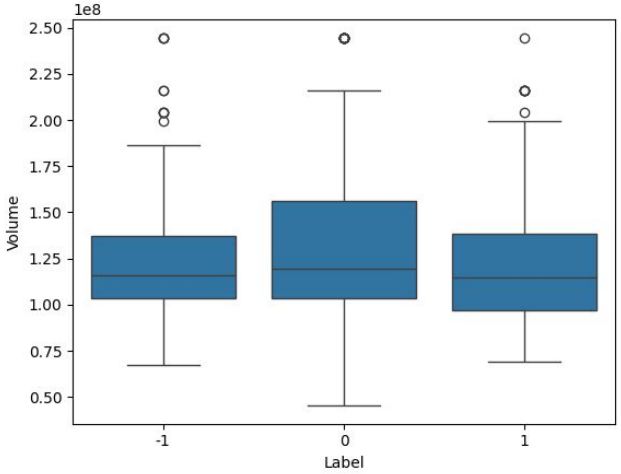


EDA Results- Bivariate Analysis

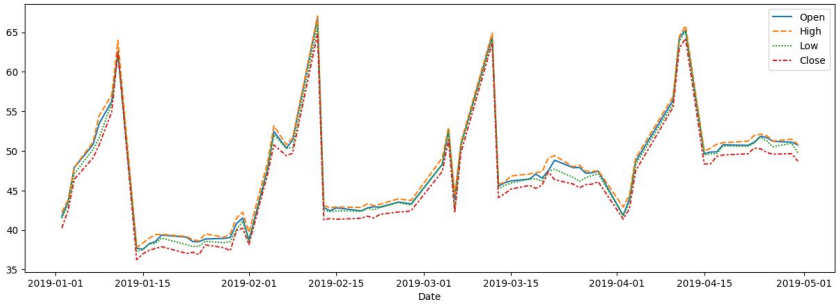
BA1- Sentiment label vs Price



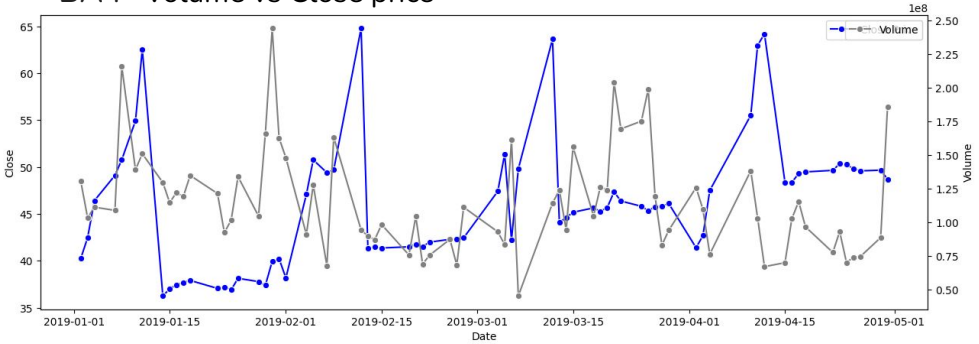
BA2- Sentiment label vs Volume



BA3- Date vs Price



BA4- Volume vs Close price



Data Background and Contents

- The dataset is composed of news articles; the shape of the dataset 349 rows by 8 columns.
- There are 4 float, 2 integer, and 2 object data types in the dataset.

Statistical summary of the dataset:

	count	mean	min	25%	50%	75%	max	std
Date	349	2019-02-16 16:05:30.085959936	2019-01-02 00:00:00	2019-01-14 00:00:00	2019-02-05 00:00:00	2019-03-22 00:00:00	2019-04-30 00:00:00	NaN
Open	349.0	46.229233	37.567501	41.740002	45.974998	50.7075	66.817497	6.442817
High	349.0	46.700458	37.817501	42.244999	46.025002	50.849998	67.0625	6.507321
Low	349.0	45.745394	37.305	41.482498	45.639999	49.7775	65.862503	6.391976
Close	349.0	44.926317	36.254131	40.246914	44.596924	49.11079	64.805229	6.398338
Volume	349.0	128948236.103152	45448000.0	103272000.0	115627200.0	151125200.0	244439200.0	43170314.918964
Label	349.0	-0.054441	-1.0	-1.0	0.0	0.0	1.0	0.715119

Data Dictionary

* **Date**: The date the news was released
* **News**: The content of news articles that could potentially affect the company's stock price
* **Open**: The stock price (in $) at the beginning of the day
* **High**: The highest stock price (in $) reached during the day
* **Low**: The lowest stock price (in $) reached during the day
* **Close**: The adjusted stock price (in $) at the end of the day
* **Volume**: The number of shares traded during the day
* **Label**: The sentiment polarity of the news content
    * 1: Positive
    * 0: Neutral
    * -1: Negative

This dataset is composed of similarly priced, actively traded stocks.

Great Learning

Happy Learning !