Stock Market News

Sentiment Analysis & Summarization
NLP & LLM

Azin Faghihi
Role: Data Scientist
May 2025

Contents / Agenda

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

Executive Summary

- We observe that the *classes* of the sentiment labels are ***imbalanced***.
- It is observed that Open, Low, Close, and High are *highly correlated*.
- We also observe that *negative sentiments* are mostly associated with *low prices* in Open, Low, Close, and High values.
- The ***length*** *of neutral news* tend to be a little shorter on average compared to the negative and positive news.
- The neutral-sentiment news on average are more associated with higher **Volume**.

Sentiment Analysis Modeling*:* We conclude that a *Tuned Random Forest* classifier trained on the *GloVe*-embedded news text lead to a model with the best trade-off between bias and overfitting with the *F1* criterion.

Content Summarization: Using an LLM model and doing prompt engineering we were about get a summary of three positive and negative news for the weekly stock news.

Business Problem Overview and Solution Approach

Our investment startup is using AI to analyze how news affects the stock price of a NASDAQ-listed company. The goal is to build a sentiment analysis system that processes daily news, summarizes it weekly, and links it to stock trends to support smarter investment decisions.

- Sentiment Analysis:
  - Using Natural Language Processing to embed the news content and use machine learning classifiers to predict the sentiment of the news content.
- Content Summarization:
  - Using LLM & Prompt Engineering to get summary of positive and negative news of the weekly news reports.

Data
Sample

- Here’s the sample of the data, the first five rows:

	Date	News	Open	High	Low	Close	Volume	Label
0	2019-01-02	The tech sector experienced a significant dec...	41.740002	42.244999	41.482498	40.246914	130672400	-1
1	2019-01-02	Apple lowered its fiscal Q1 revenue guidance ...	41.740002	42.244999	41.482498	40.246914	130672400	-1
2	2019-01-02	Apple cut its fiscal first quarter revenue fo...	41.740002	42.244999	41.482498	40.246914	130672400	-1
3	2019-01-02	This news article reports that yields on long...	41.740002	42.244999	41.482498	40.246914	130672400	-1
4	2019-01-02	Apple's revenue warning led to a decline in U...	41.740002	42.244999	41.482498	40.246914	130672400	-1

Data

- There are 349 rows and 8 columns in the original dataset. One column (News Length) is added. The Volume value is also modified to be reported in millions.
- The ***memory usage*** is approximately 24.7 KB.
- There are ***no missing values*** in the data.
- There are ***no duplicated rows*** in the data.

Memory Usage	24.7 KB

#	
Rows	349
Columns	9
Null Values	0
Duplicated Rows	0

Data Dictionary

Column	Data Type	Description	# unique
Date	datetime64[ns]	The date the news was released	71
News	object	The content of news articles that could potentially affect the company's stock price	349
Open	float64	The stock price (in $) at the beginning of the day	70
High	float64	The highest stock price (in $) reached during the day	70
Low	float64	The lowest stock price (in $) reached during the day	71
Close	float64	The adjusted stock price (in $) at the end of the day	71
Volume (M)	float64	The number of shares traded during the day (in millions)	71
Label	int64	The sentiment polarity of the news content, 1.positive; 0.neutral, -1.negative	3
News Length	int64	Length of the news text	30

Data

- The following tables show the summary information of our variables.
  - ***Categorical***: unique counts, most common value and its corresponding frequency
  - ***Numerical***: mean, median, standard deviation, minimum, maximum, outlier counts, ...

Object/Categorical Column	unique	top	freq
Label	3	0	170

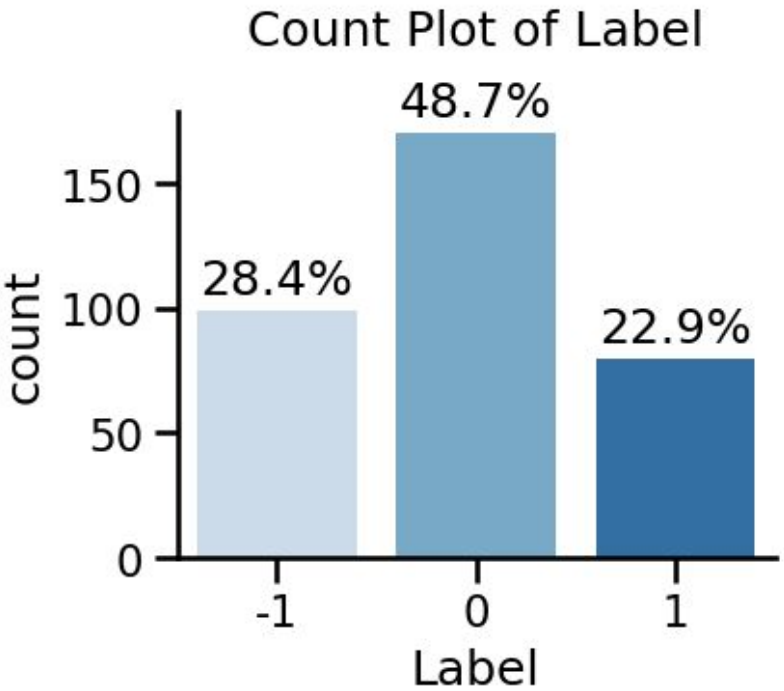
	count	mean	min	25%	50%	75%	max
Date	349	2019-02-16 16:05:30.085959936	2019-01-02 00:00:00	2019-01-14 00:00:00	2019-02-05 00:00:00	2019-03-22 00:00:00	2019-04-30 00:00:00

Numerical Column	mean	std	min	25%	50%	75%	max	IQR	# Outliers (Upper)	# Outliers (Lower)	# Outliers	Outliers %
Open	46.2	6.4	37.5675	41.74	45.975	50.7075	66.8175	8.9675	11	0	11	3.2
High	46.7	6.5	37.8175	42.245	46.025	50.85	67.0625	8.605	17	0	17	4.9
Low	45.7	6.4	37.305	41.4825	45.64	49.7775	65.8625	8.295	17	0	17	4.9
Close	44.9	6.4	36.2541	40.2469	44.5969	49.1108	64.8052	8.86388	17	0	17	4.9
Volume (M)	128.9	43.2	45.448	103.272	115.627	151.125	244.439	47.8532	17	0	17	4.9
News Length	49.3	5.7	19	46	50	53	61	7	0	6	6	1.7

EDA - Univariate, Categorical

Label

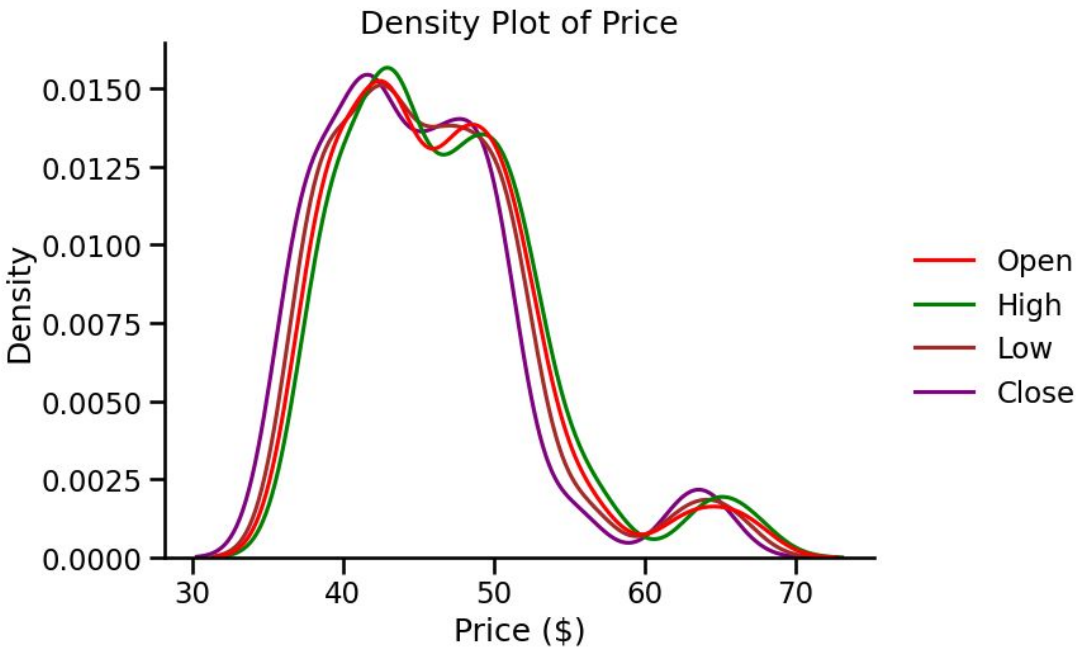
- The most common sentiment is neutral (~49 %).



EDA - Numerical

Open, High, Low, Close

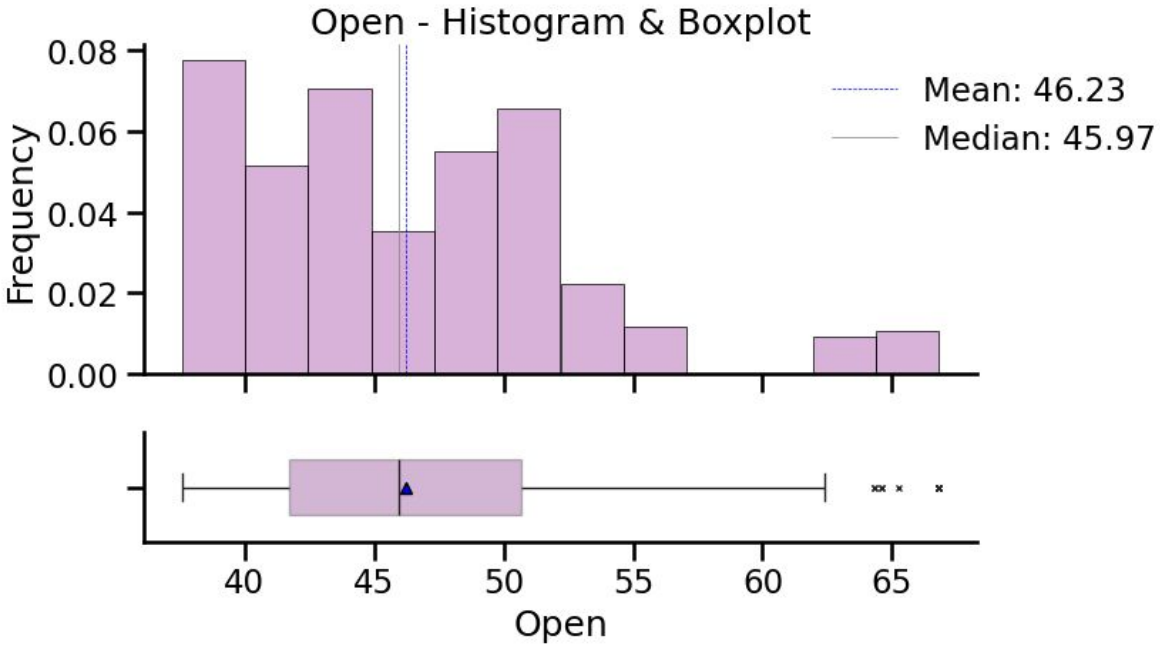
- Here’s the density plot of price values for Open, High, Low and Close.



EDA - Univariate, Numerical

Open

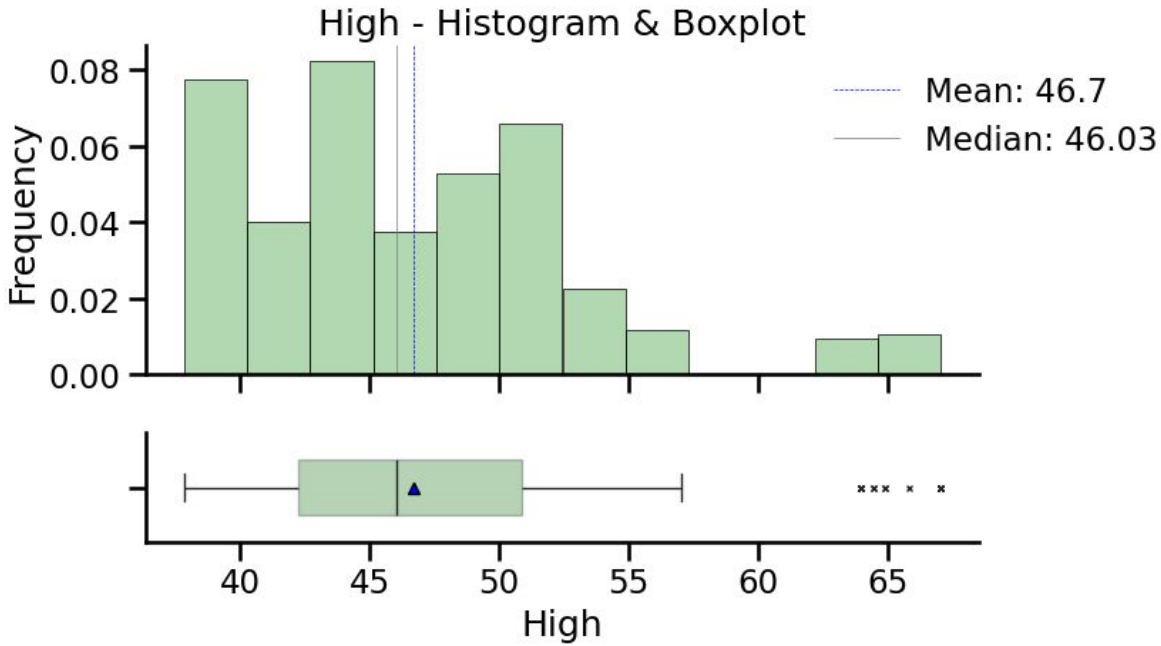
- Open is highly positively skewed (right-skewed) (skewness: 1.01) with average of 46.2.



EDA - Univariate, Numerical

High

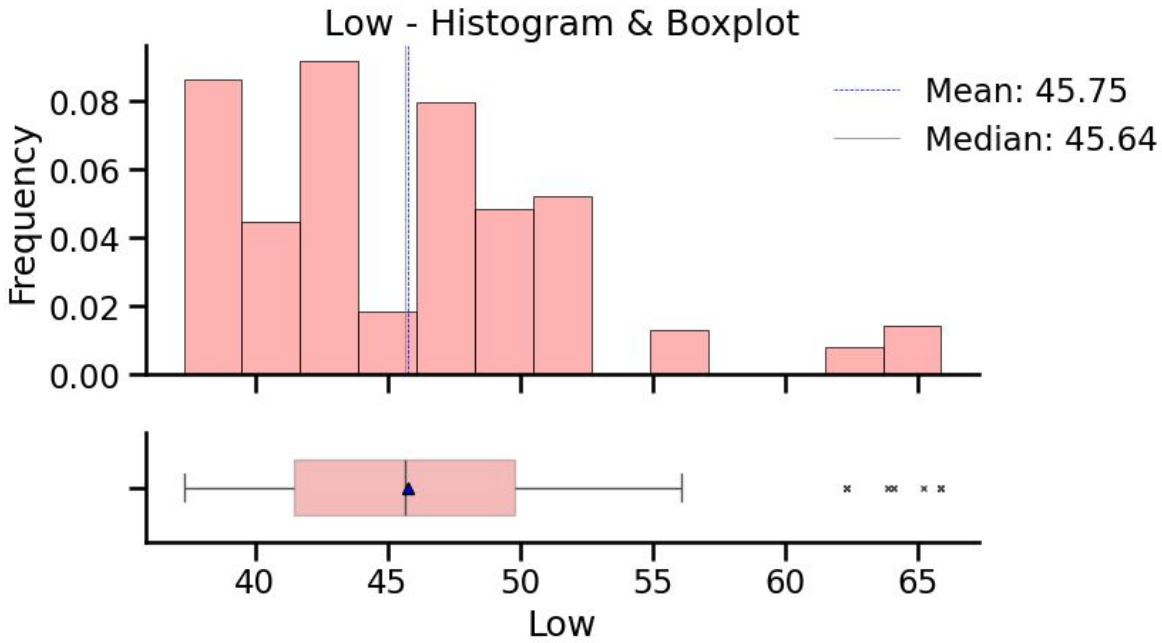
- High is highly positively skewed (right-skewed) (skewness: 1.02) with average of 46.7.



EDA - Univariate, Numerical

Low

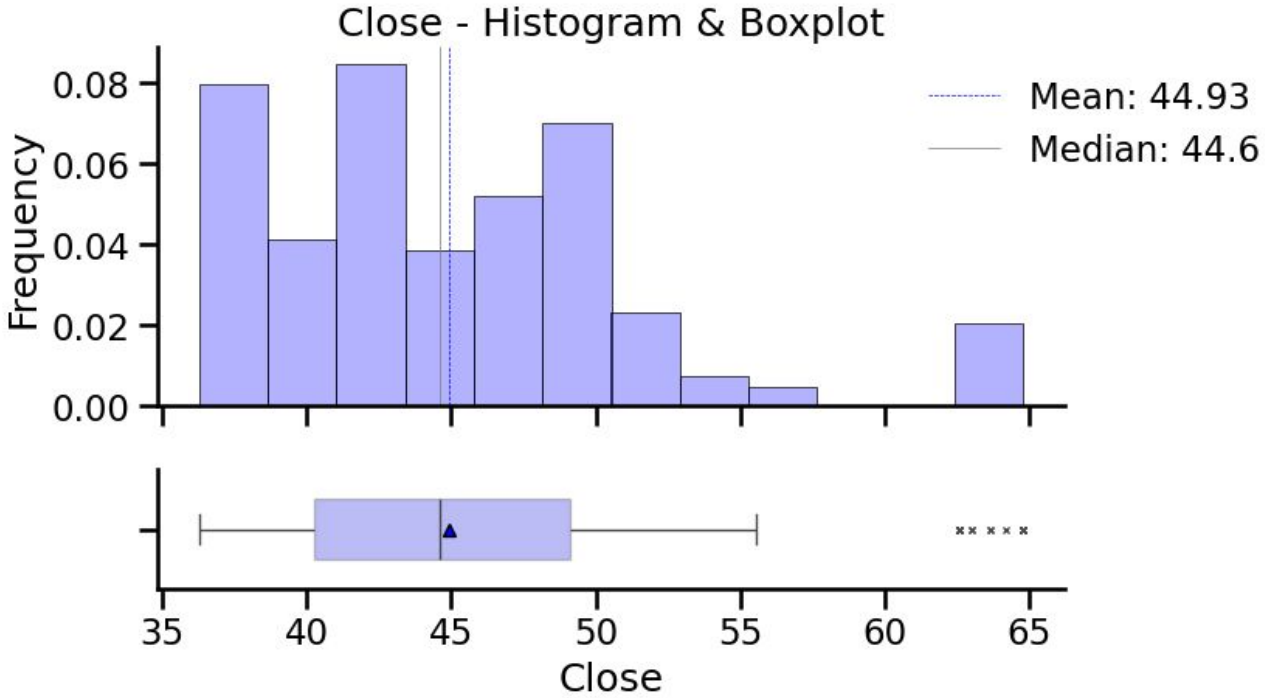
- Low is highly positively skewed (right-skewed) (skewness: 1.02) with average of 45.7.



EDA - Univariate, Numerical

Close

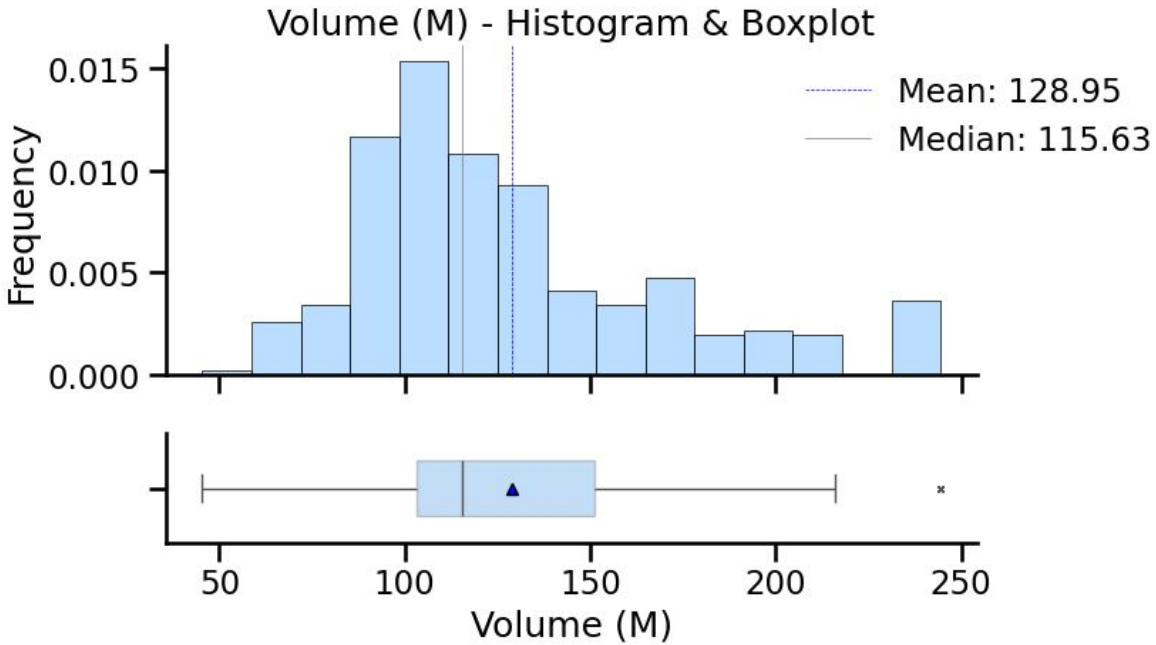
- Close is highly positively skewed (right-skewed) (skewness: 1.06) with average of 44.9.



EDA - Univariate, Numerical

Volume

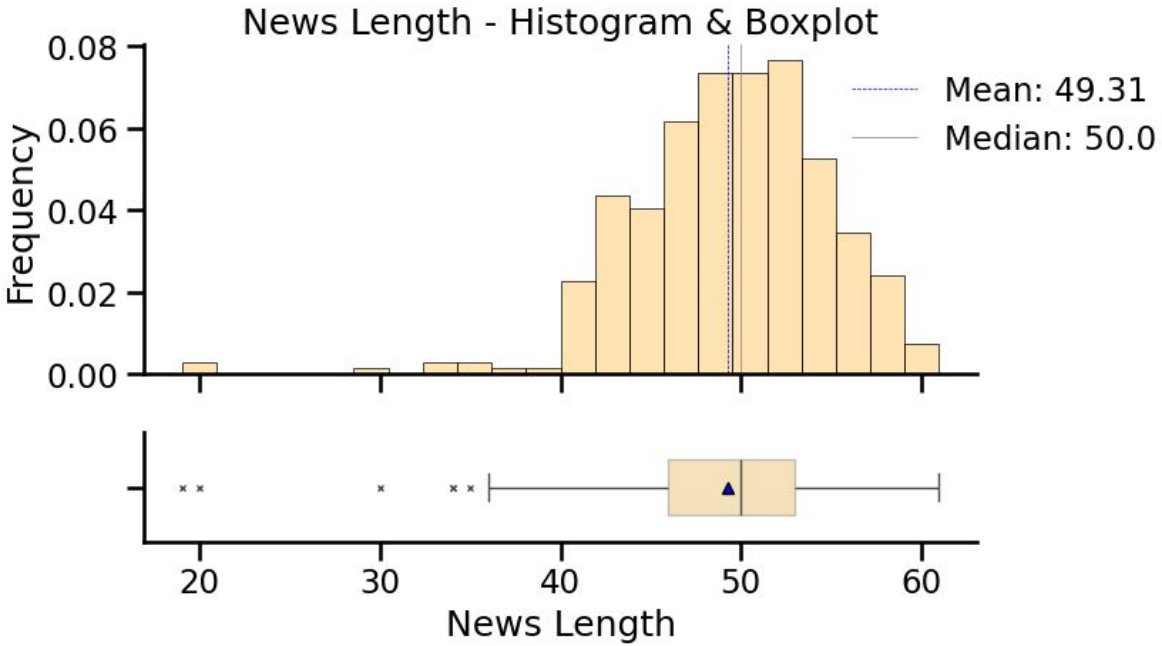
- Volume (M) is highly positively skewed (right-skewed) (skewness: 1.11) with average of 128.95 million.



EDA - Univariate, Numerical

News Length

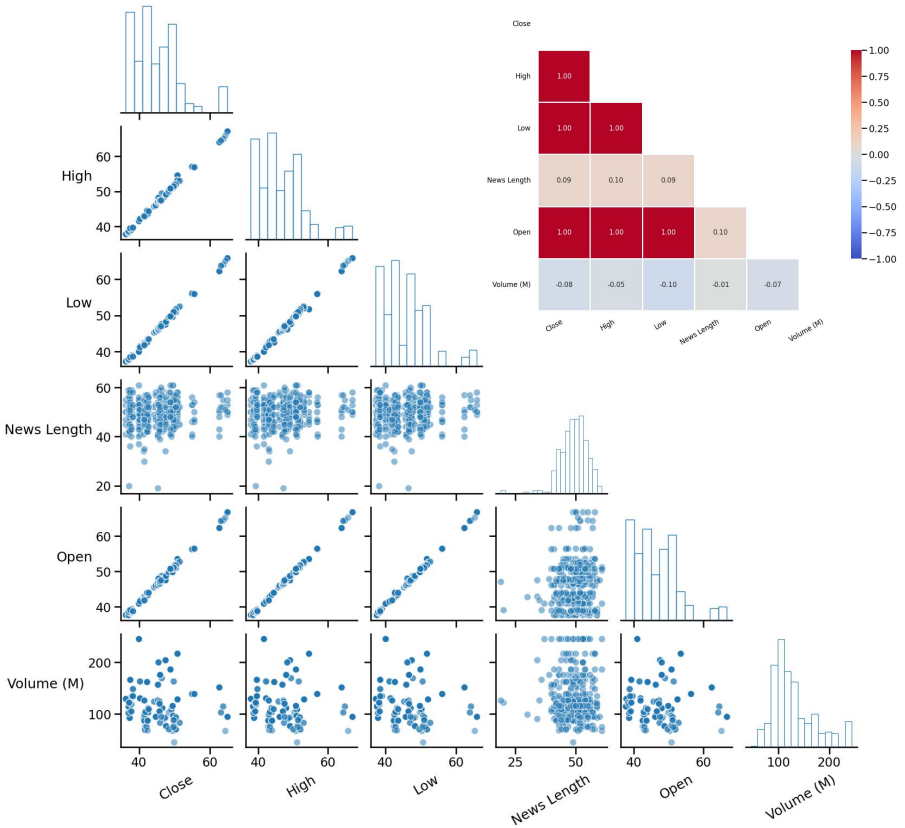
- News Length is highly negatively skewed (left-skewed) (skewness: -0.98) with average of 49 words.



EDA - Bivariate, Numerical

- There's a high correlation between High, Open, Low, and Close.

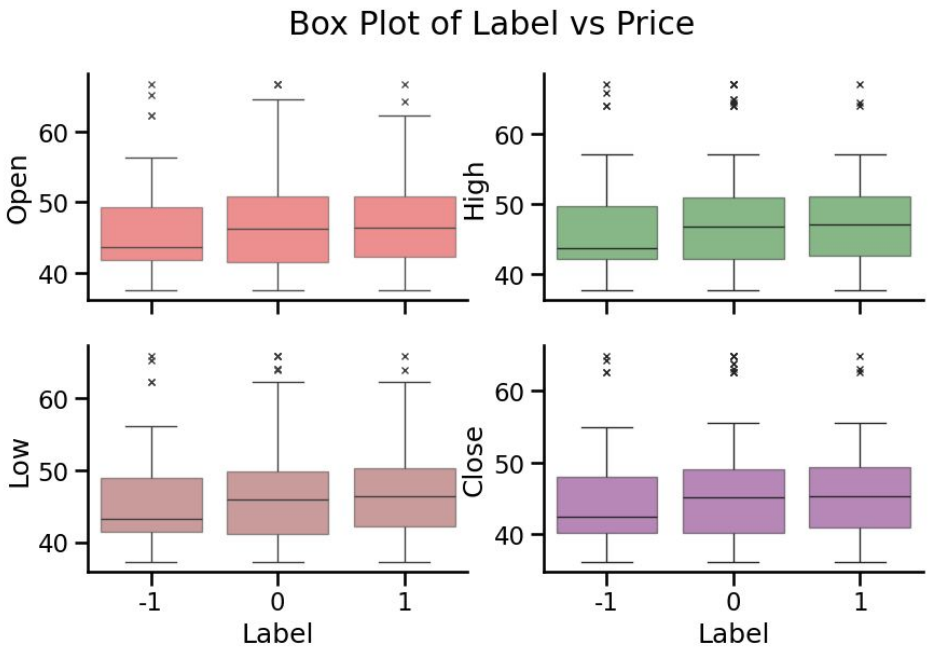
Variable 1	Variable 2	Correlation
High	Open	0.998526
Close	Low	0.998453
Low	Open	0.997900
Close	High	0.997501
High	Low	0.997328
Close	Open	0.996273



EDA - Bivariate

Label vs Price

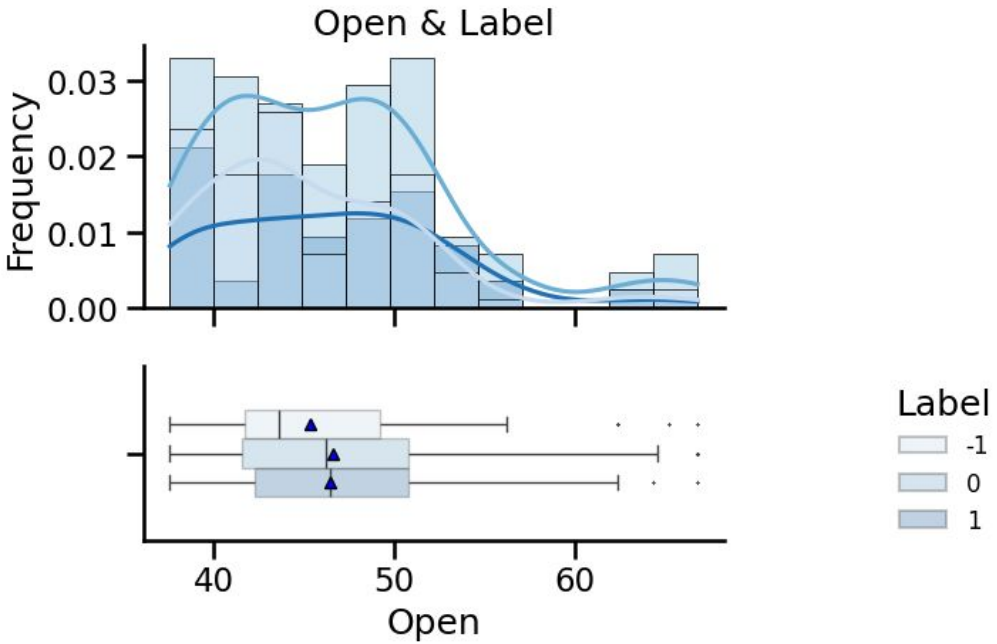
- The price values on average seems to be lower amongst the negative-sentiment news.



EDA - Bivariate

Open vs Label

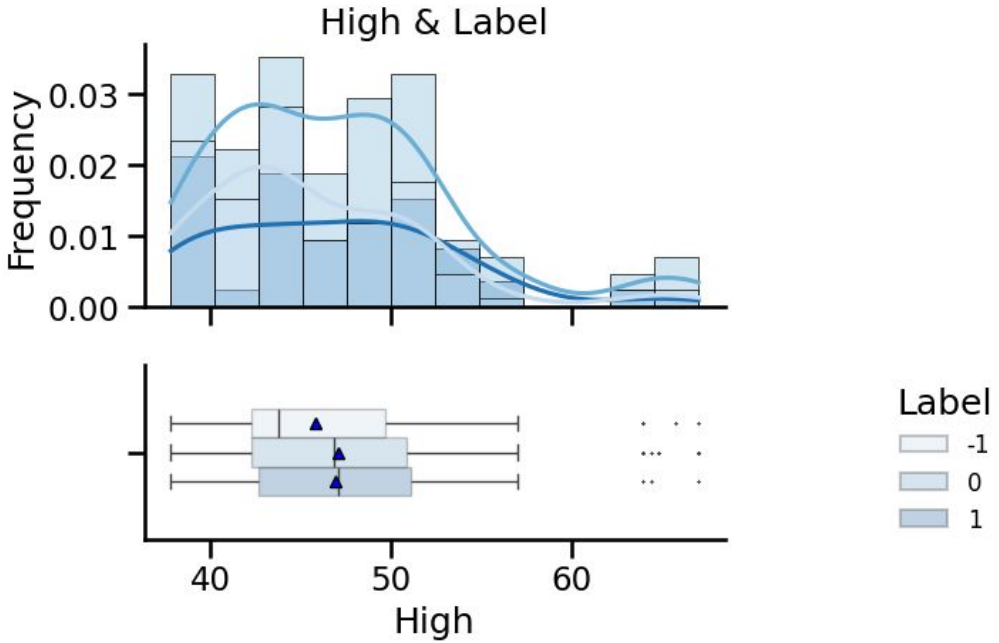
- The Open value on average seems to be lower amongst the negative-sentiment news.



EDA - Bivariate

High vs Label

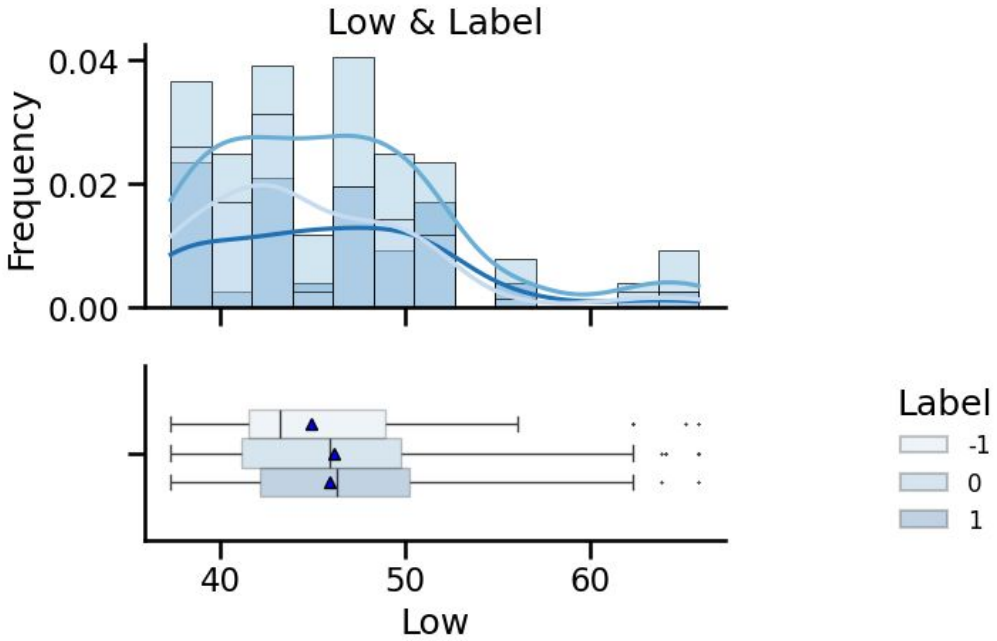
- The High value on average seems to be lower amongst the negative-sentiment news.



EDA - Bivariate

Low vs Label

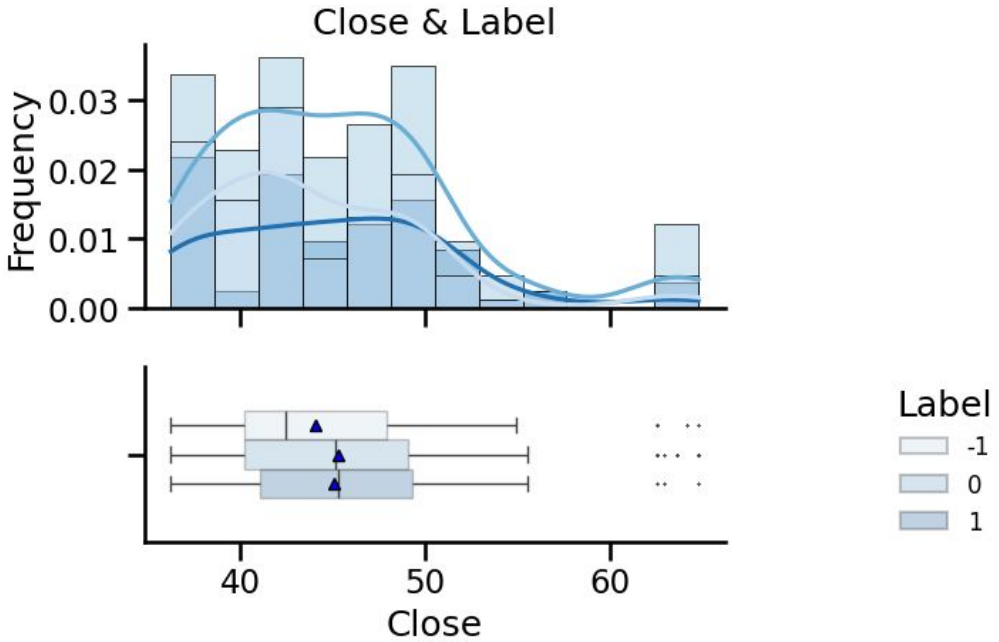
- The Low value on average seems to be lower amongst the negative-sentiment news.



EDA - Bivariate

Close vs Label

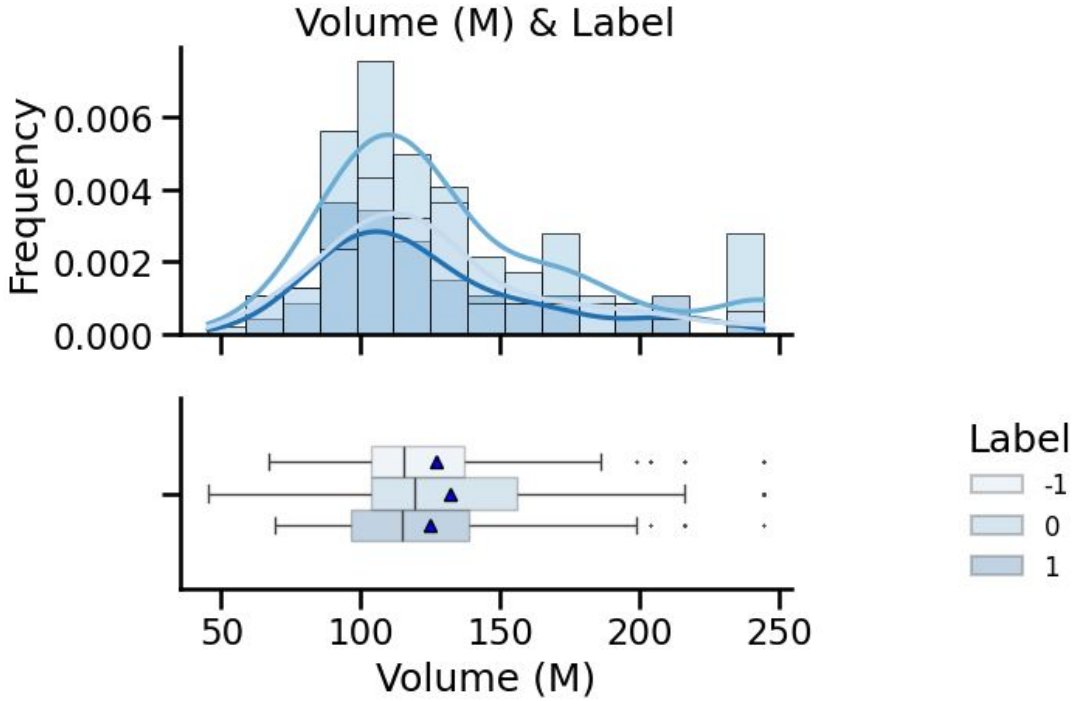
- The Close value on average seems to be lower amongst the negative-sentiment news.



EDA - Bivariate

Volume vs Label

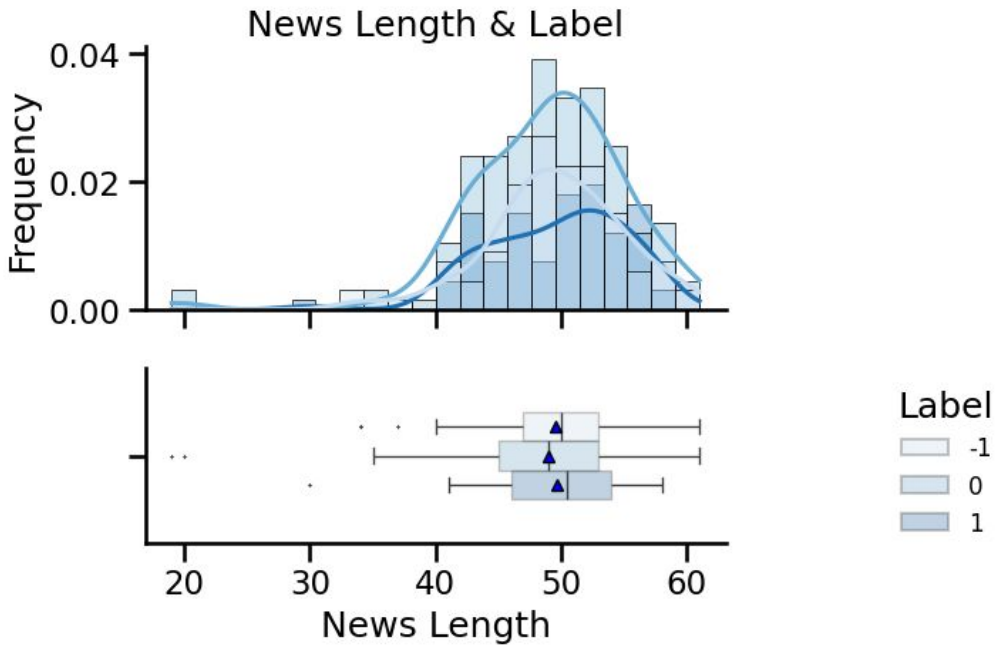
- The volume corresponding to the neutral-sentiment news on average seems to be higher.



EDA - Bivariate

News Length vs Label

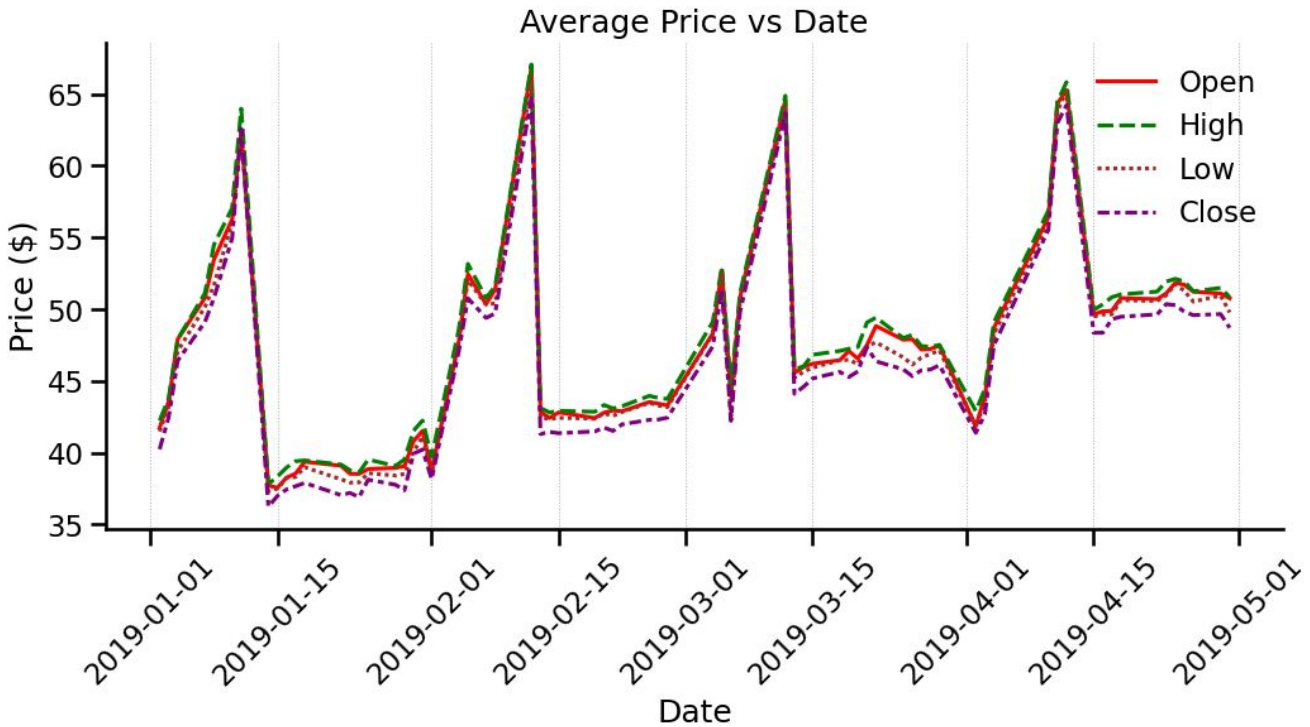
- The length of the neutral-sentiment news on average seems to be less than the rest.



EDA - Multivariate

Avg Price vs Date

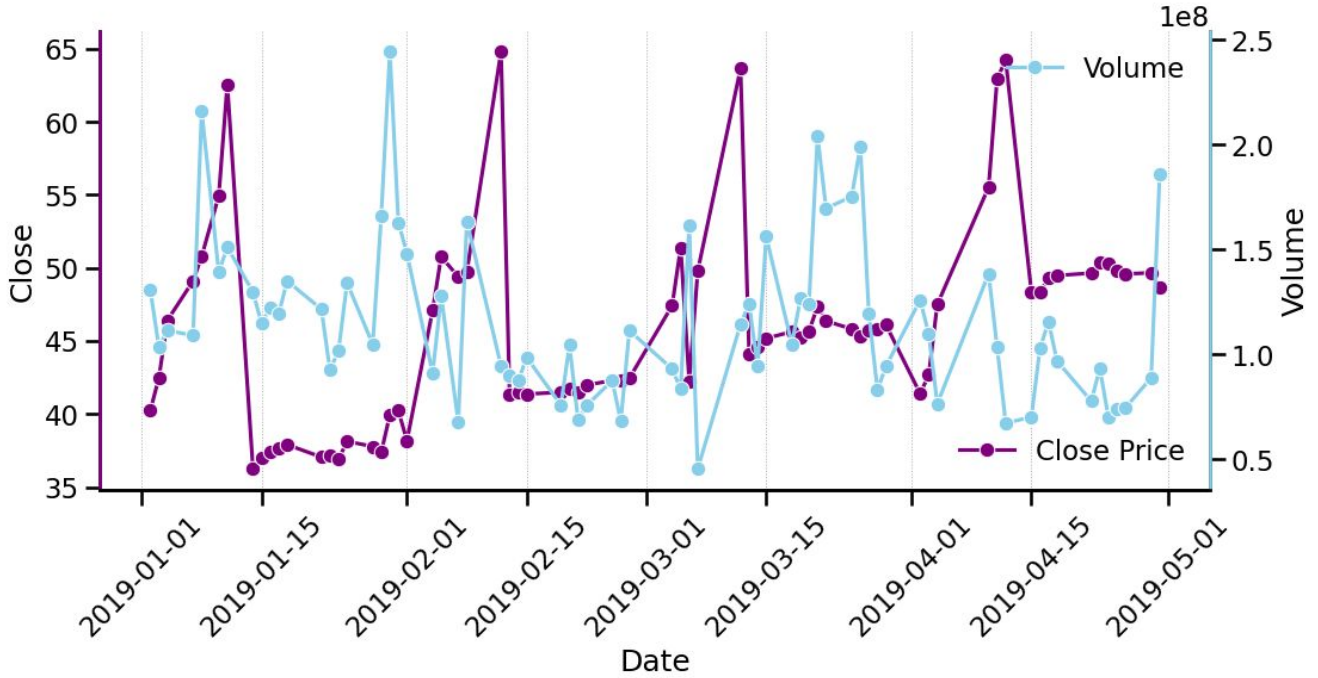
- Here we see the average price vs date values.



EDA - Multivariate

Avg Close & Volume vs Date

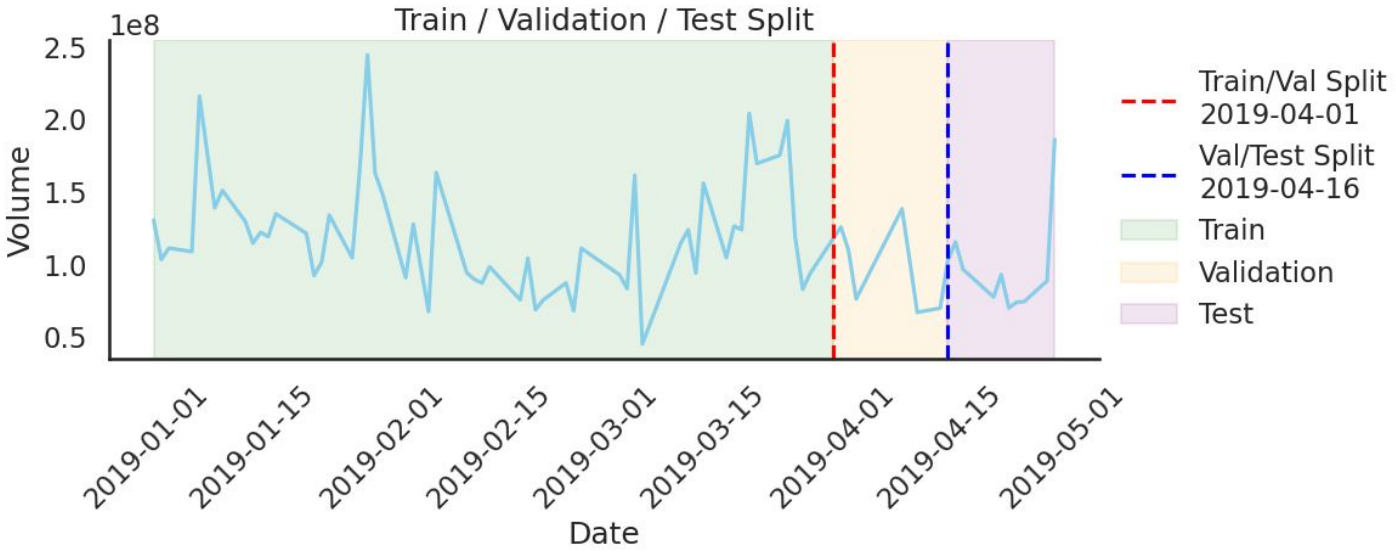
- Here we see the close and volume trends vs date.



Data Preprocessing

Train/Validation/Test Split

- The data is split into train, validation and test sets on select dates as follows:



Shape	X	y
Train	286 x 10	286
Validation	21 x 10	21
Test	42 x 10	42

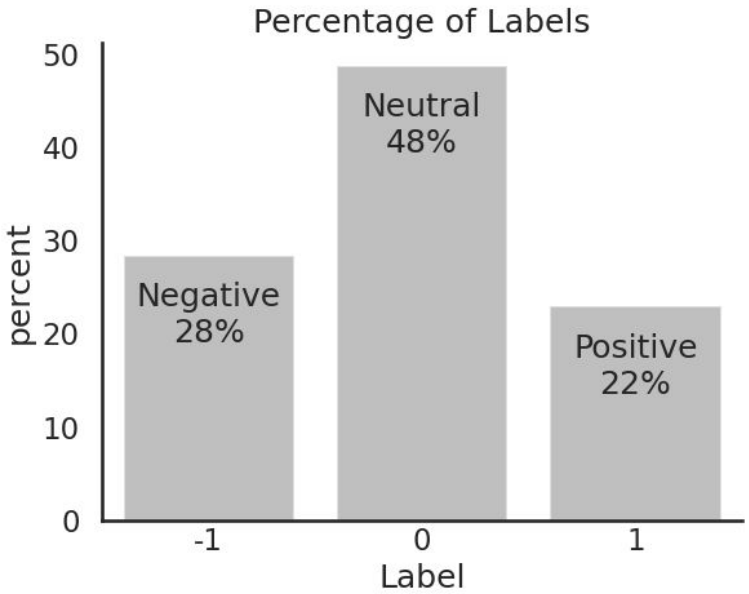
Data Preprocessing
Word2Vec, GloVe, Sentence Transformer

- We use three types of embeddings to convert the raw text in the ‘News’ into numerical vectors to be used by our machine learning models.
  - Word2Vec : size: 300, min count: 1, window: 5, workers: 6
  - GloVe: size: 100
  - Sentence Transformer: size: 384

X	Original	Word2Vec	GloVe	Sentence Transformer
Train	286 x 10	286 x 300	286 x 100	286 x 384
Validation	21 x 10	21 x 300	21 x 100	21 x 384
Test	42 x 10	42 x 300	42 x 100	42 x 384

Sentiment Analysis - Model Evaluation Criterion

- Since the classes are **imbalanced**, we choose **F1** as our model evaluation criterion.



Sentiment Analysis - Model Building

Base Model Performance

- Following are the performance metrics for the base models on the three embeddings.
- All these models seem to **overfit** the data since the |training-validation| is significant.

Base Model Performance	Accuracy	Recall	Precision	F1
Train - Word2Vec I GradientBoosting	1.000000	1.000000	1.000000	1.000000
Validation - Word2Vec I GradientBoosting	0.428571	0.428571	0.387755	0.402597
Train - GloVe I RandomForest	1.000000	1.000000	1.000000	1.000000
Validation - GloVe I RandomForest	0.476190	0.476190	0.400794	0.426871
Train - Sentence Transformer I DecisionTree	1.000000	1.000000	1.000000	1.000000
Validation - Sentence Transformer I DecisionTree	0.523810	0.523810	0.443537	0.480260

Sentiment Analysis - Model Improvement

- In order to improve the model performance and reduce the overfitting, we perform hyperparameter tuning.
- For each base model, we select a parameter grid and selected the best performing model using grid search. (max depth, min samples split, max features)
- Here are the performance metrics for the tuned models.

Tuned Model Performance	Accuracy	Recall	Precision	F1
Train - Word2Vec I GradientBoosting	1.000000	1.000000	1.000000	1.000000
Validation - Word2Vec I GradientBoosting	0.380952	0.380952	0.314286	0.343915
Train - GloVe I RandomForest	0.979021	0.979021	0.979545	0.978968
Validation - GloVe I RandomForest	0.571429	0.571429	0.539683	0.530612
Train - Sentence Transformer I DecisionTree	0.702797	0.702797	0.727462	0.703814
Validation - Sentence Transformer I DecisionTree	0.285714	0.285714	0.265306	0.272727

Sentiment Analysis – Model Performance Comparison

- Here we have the summary of key performance metrics for training and validation data of all the models (base+tuned) for comparison
- We also included the difference between the train and validation scores for our criterion F1

Training Performance Comparison	Accuracy	Recall	Precision	F1
Word2Vec I GradientBoosting	1.000000	1.000000	1.000000	1.000000
GloVe I RandomForest	1.000000	1.000000	1.000000	1.000000
Sentence Transformer I DecisionTree	1.000000	1.000000	1.000000	1.000000
Word2Vec I GradientBoosting I Tuned	1.000000	1.000000	1.000000	1.000000
GloVe I RandomForest I Tuned	0.979021	0.979021	0.979545	0.978968
Sentence Transformer I DecisionTree I Tuned	0.702797	0.702797	0.727462	0.703814

Validation Performance Comparison	Accuracy	Recall	Precision	F1
Word2Vec I GradientBoosting	0.428571	0.428571	0.387755	0.402597
GloVe I RandomForest	0.476190	0.476190	0.400794	0.426871
Sentence Transformer I DecisionTree	0.523810	0.523810	0.443537	0.480260
Word2Vec I GradientBoosting I Tuned	0.380952	0.380952	0.314286	0.343915
GloVe I RandomForest I Tuned	0.571429	0.571429	0.539683	0.530612
Sentence Transformer I DecisionTree I Tuned	0.285714	0.285714	0.265306	0.272727

|Train – Validation| Performance Difference	F1
Sentence Transformer I DecisionTree I Tuned	0.431087
GloVe I RandomForest I Tuned	0.448356
Sentence Transformer I DecisionTree	0.519740
Word2Vec I GradientBoosting	0.572161
GloVe I RandomForest	0.573129
Word2Vec I GradientBoosting I Tuned	0.685714

Sentiment Analysis – Final Model

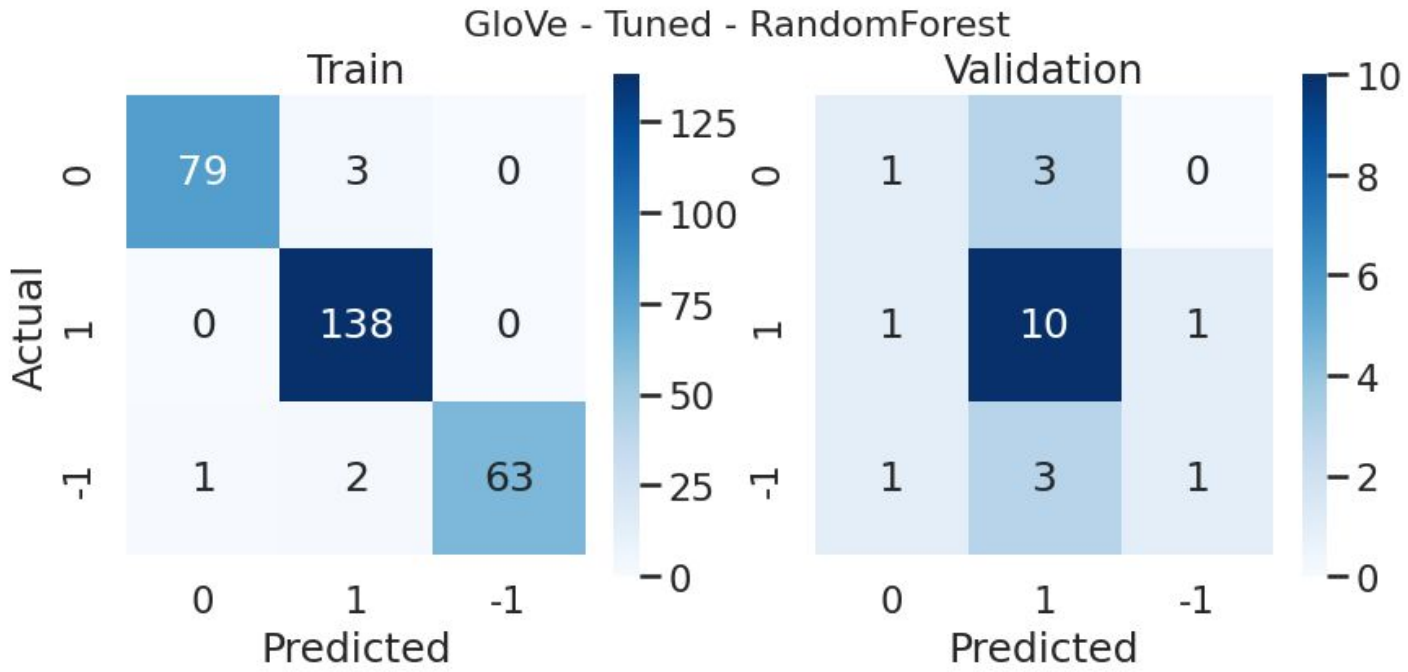
- For our final model, we would like to choose a model with low bias that does not also overfit.
- We look at the score difference between training and validation and would like to choose the the models with low score difference which would give us a model that overfits the least amongst the other models.
- We pick the model that has the highest training and validation scores with the lowest overfitting: Glove embedding, Tuned Random Forest model

	F1
Training Performance Comparison	
GloVe I RandomForest I Tuned	0.978968
Sentence Transformer I DecisionTree I Tuned	0.703814
Validation Performance Comparison	
GloVe I RandomForest I Tuned	0.530612
Sentence Transformer I DecisionTree	0.480260

|Train - Validation| Performance Difference	F1
Sentence Transformer I DecisionTree I Tuned	0.431087
GloVe I RandomForest I Tuned	0.448356
Sentence Transformer I DecisionTree	0.519740
Word2Vec I GradientBoosting	0.572161
GloVe I RandomForest	0.573129
Word2Vec I GradientBoosting I Tuned	0.685714

Sentiment Analysis – Final Model

- Here are the confusion matrices for the final model in training and validation

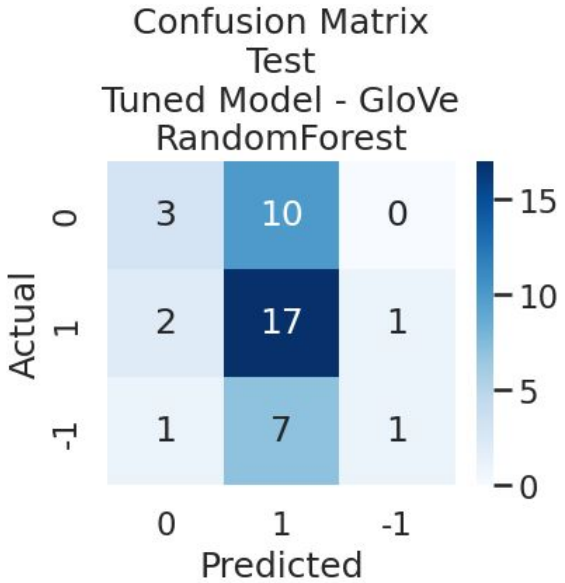


Sentiment Analysis – Final Model

- Here's the performance of our model on the test data.
- Provided are also the model parameters for our final model.

Final Model Performance (Test)	Accuracy	Recall	Precision	F1
Test - GloVe I RandomForest I Tuned Model	0.5	0.5	0.5	0.436529

Final Model Parameters – RandomForest	Value
bootstrap	True
criterion	gini
max_depth	6
max_features	0.2
min_samples_leaf	1
min_samples_split	7
n_estimators	100
random_state	42



Content Summarization – Data Preprocessing

- The original data as discussed before is of the shape 349 x 8
- For this step, we choose the **Date** & **News** columns and **aggregate** the records on a **weekly basis** instead.
- Shape of our data is now 18 x 2

	Date	News
0	2019-01-06	The tech sector experienced a significant dec...
1	2019-01-13	Sprint and Samsung plan to release 5G smartph...
2	2019-01-20	The U.S. stock market declined on Monday as c...
3	2019-01-27	The Swiss National Bank (SNB) governor, Andre...
4	2019-02-03	Caterpillar Inc reported lower-than-expected ...
5	2019-02-10	The Dow Jones Industrial Average, S&P 500, an...

Content Summarization – Modeling Approach

- We use Large Language Model (LLM) LLAMA from llama_cpp package.

LLM Parameters:

- **Model Name**: "TheBloke/Mistral-7B-Instruct-v0.2-GGUF"
- **Number of GPU Layers**: 100
- **Context Window**: 4500
- **Maximum Number of Token**: 1024
- **Temperature**: 0
- **Top p**: .95
- **Top k**: 50

Content Summarization – Sample Input/Output

Here are the sample input and the prompt used for this task:

Sample Input

```
 The tech sector experienced a significant decline in the
aftermarket following Apple's Q1 revenue warning. Notable
suppliers, including Skyworks, Broadcom, Lumentum, Qorvo,
and TSMC, saw their stocks drop in response to Apple's
downward revision of its revenue expectations for the
quarter, previously announced in January. ||  Apple lowered
its fiscal Q1 revenue guidance to $84 billion from earlier
estimates of $89–$93 billion due to weaker than expected
iPhone sales. The announcement caused a significant drop in
Apple's stock price and negatively impacted related
suppliers, leading to broader market declines for tech
indices such as Nasdaq 10 ||  Apple cut its fiscal first
quarter revenue forecast from $89–$93 billion to $84 billion
due to weaker demand in China and fewer iPhone upgrades. CEO
Tim Cook also mentioned constrained sales of Airpods and
Macbooks. Apple's shares fell 8.5% in post market trading,
while Asian suppliers like Hon ||  This news article reports
that yields on long-dated U.S. Treasury securities hit their
lowest levels in nearly a year on January 2, 2019, due to
concerns about the health of the global economy following
weak economic data from China and Europe, ...
```

Prompt

```
You are an expert data analyst specializing in stock-related news content analysis.

Task: Analyze the provided stock news and identify the top three positive and negative events
that are most likely to impact the price of the stock.

Instructions:
1. Read and analyze the stock news content.
2. Identify the three most positive events.
3. Identify the three most negative events.
4. Summarize each event concisely.

Return the output in JSON format containing two keys, Positive Events and Negative Events.
The values are list of events as in the following structure:

{
  "Positive Events": [
    "First top positive event",
    "Second top positive event",
    "Third top positive event"
  ],
  "Negative Events": [
    "First top negative event",
    "Second top negative event",
    "Third top negative event"
  ]
}
```

Content Summarization – Sample Input/Output

- Below is the sample output.

```
{
    "Positive Events": [
      "Roku Inc announced plans to offer premium video channels on a subscription
      basis through its free streaming service, The Roku Channel.",
      "The Supreme Court will review Broadcom's appeal in a shareholder lawsuit
      over the 2015 acquisition of Emulex.",
      "The Chinese central bank announced a fifth reduction in the required reserve
      ratio (RRR) for banks, freeing up approximately 116.5 billion yuan for new lending."
    ],
    "Negative Events": [
      "Apple cut its fiscal first quarter revenue forecast from $89-$93 billion
      to $84 billion due to weaker demand in China and fewer iPhone upgrades.",
      "Apple's revenue warning led to a decline in USD JPY pair and a gain in
      Japanese yen, as investors sought safety in the highly liquid currency.",
      "Apple CEO Tim Cook discussed the company's Q1 warning on CNBC, attributing
      US-China trade tensions as a factor."
    ]
}
```

Content Summarization – Raw Model Output

- Here’s the snapshot of the resultant dataframe for the first three rows.

	Date	News	Key Events	model_response_parsed
0	2019-01-06	The tech sector experienced a significant decline in th...	{\n "Positive Events": [\n "Roku In...	{'Positive Events': ['Roku Inc announced plans to offer ...
1	2019-01-13	Sprint and Samsung plan to release 5G smartphones in ni...	{\n "Positive Events": [\n "Sprint ...	{'Positive Events': ['Sprint and Samsung's plan to relea...
2	2019-01-20	The U.S. stock market declined on Monday as concerns ov...	{\n "Positive Events": [\n "Dialog ...	{'Positive Events': ['Dialog Semiconductor reported resi...

Content Summarization – Final Output

- Below is the snapshot of the final dataframe.

	Week End Date	News	Week Positive Events	Week Negative Events
0	2019-01-06 00:00:00	The tech sector experienced a significant decline in the aftermarket following Apple's Q1 revenue warning. Notable suppliers, including Skyworks, Broadcom, Lumentum, Qorvo, and TS...	[Roku Inc announced plans to offer premium video channels on a subscription basis through its free streaming service, The Roku Channel., The Supreme Court will review Broadcom's app...]	[Apple cut its fiscal first quarter revenue forecast from $89-$93 billion to $84 billion due to weaker demand in China and fewer iPhone upgrades., Apple's revenue warning led to a d...]
1	2019-01-13 00:00:00	Sprint and Samsung plan to release 5G smartphones in nine U.S. cities this summer, with Atlanta, Chicago, Dallas, Houston, Kansas City, Los Angeles, New York City, Phoenix, and Wa...	[Sprint and Samsung's plan to release 5G smartphones in nine U.S. cities this summer., AMS developing a new 3D facial recognition sensor that can be placed behind a smartphone's scr...]	[Geely forecasting flat sales for 2019 due to economic slowdown and cautious consumers., Chinese smartphone market experiencing a decline of 12-15.5 percent in shipments last year w...]
2	2019-01-20 00:00:00	The U.S. stock market declined on Monday as concerns over a global economic slowdown intensified following unexpected drops in China's exports and imports, with tech stocks suffer...	[Dialog Semiconductor reported resilient fourth quarter revenue despite a decrease in iPhone sales at main customer Apple, leading to a 4% increase in the company's shares., Netflix...]	[China's unexpected drops in exports and imports led to a halt in Europe's four-day stock market rally, causing significant losses for technology and luxury goods sectors., The Chin...]

APPENDIX

Great Learning

Happy Learning !