Stock Market News Sentiment Analysis and Summarization

Natural Language Processing

14th April 2025

Contents / Agenda

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

Executive Summary

- Data set is carefully validated for univariant and bi variant Analysis . Data is processes for validating under training and Test ,along with validation.
- With the data, model is generated on Word2vec, Glove and sentence transformer .
- Both base and Tuned are generated for the analysis.
- Sentimental analysis is performed on the six-model generated. This will help to generate the summarization on the words based on the requirement .
- Then the final output is generated to sort the data set based on the positive and negative feedbacks

Business Problem Overview and Solution Approach

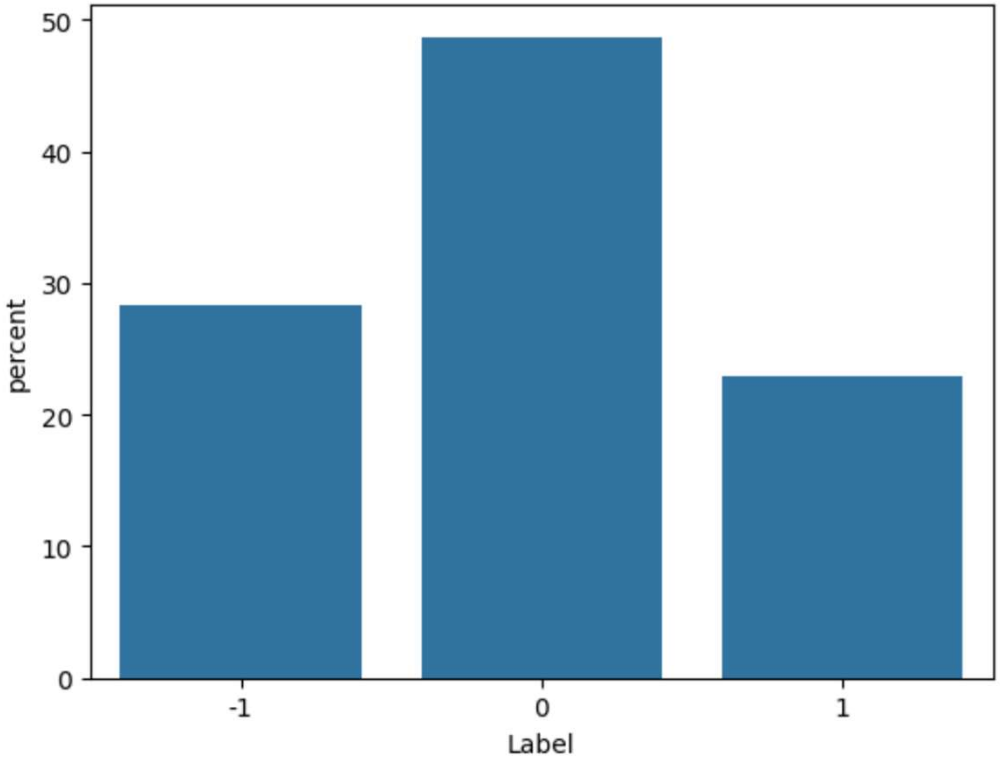
Stock price varies in every sec in this modern world. As a result , invertment firms need sophisticated tools to analyze market sentiment and integrate this information into their investment stratergies

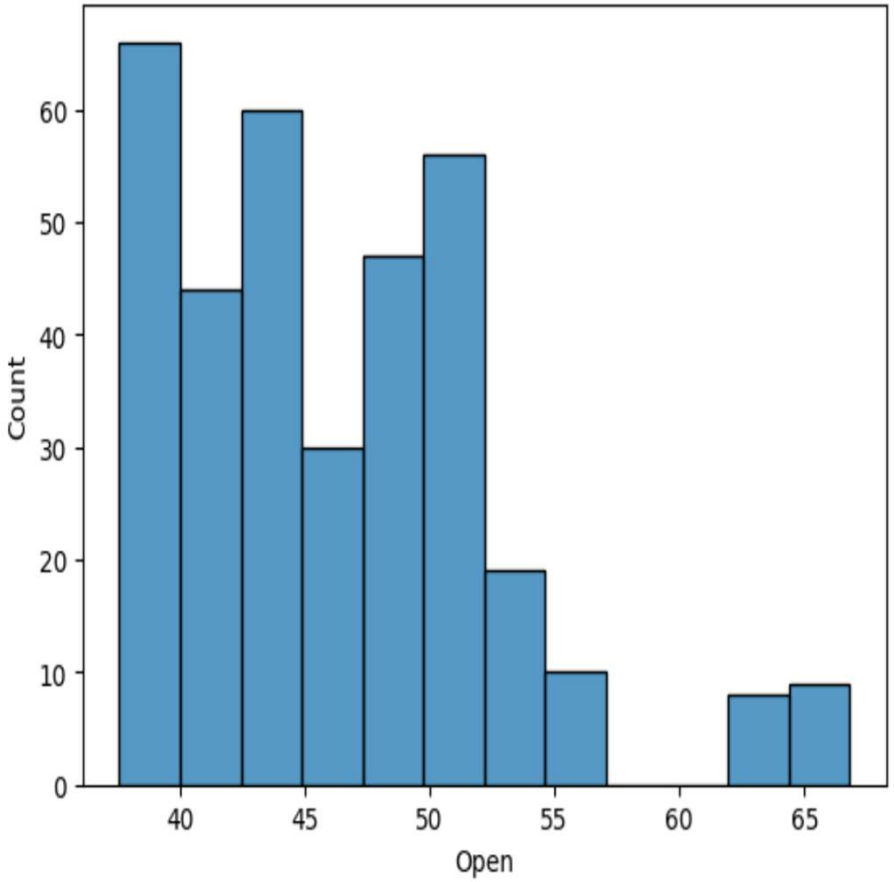
The solution approach is an AI driven sentiment analysis system that will automatically

Process and analyze news articles to guage market sentiment and summerizing the news at a news at a weekly level to enchancethe accuracy of their stock price predictions and optimize investment stratergies.

EDA Results

- In the Univariate analysis





EDA Results

- In the Univariate analysis

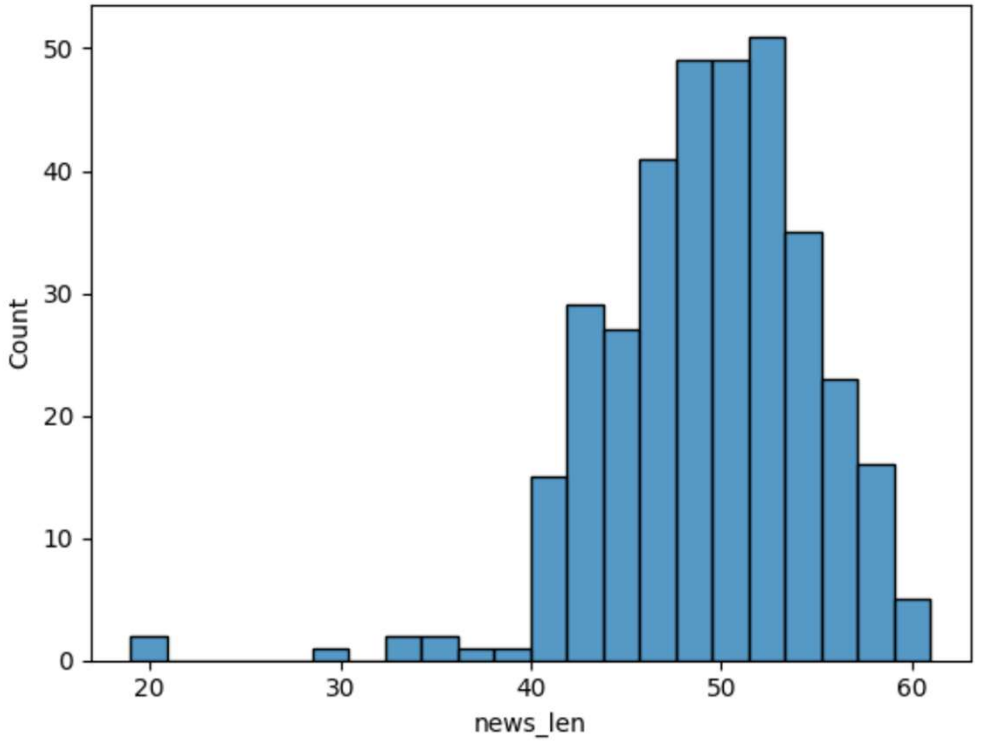
	News
0	324
1	323
2	296
3	300
4	305
...	...
344	299
345	274
346	315
347	383
348	314

349 rows × 1 columns

The univariate chart is generated to understand the behavior of one variable with respect to dependent variable.

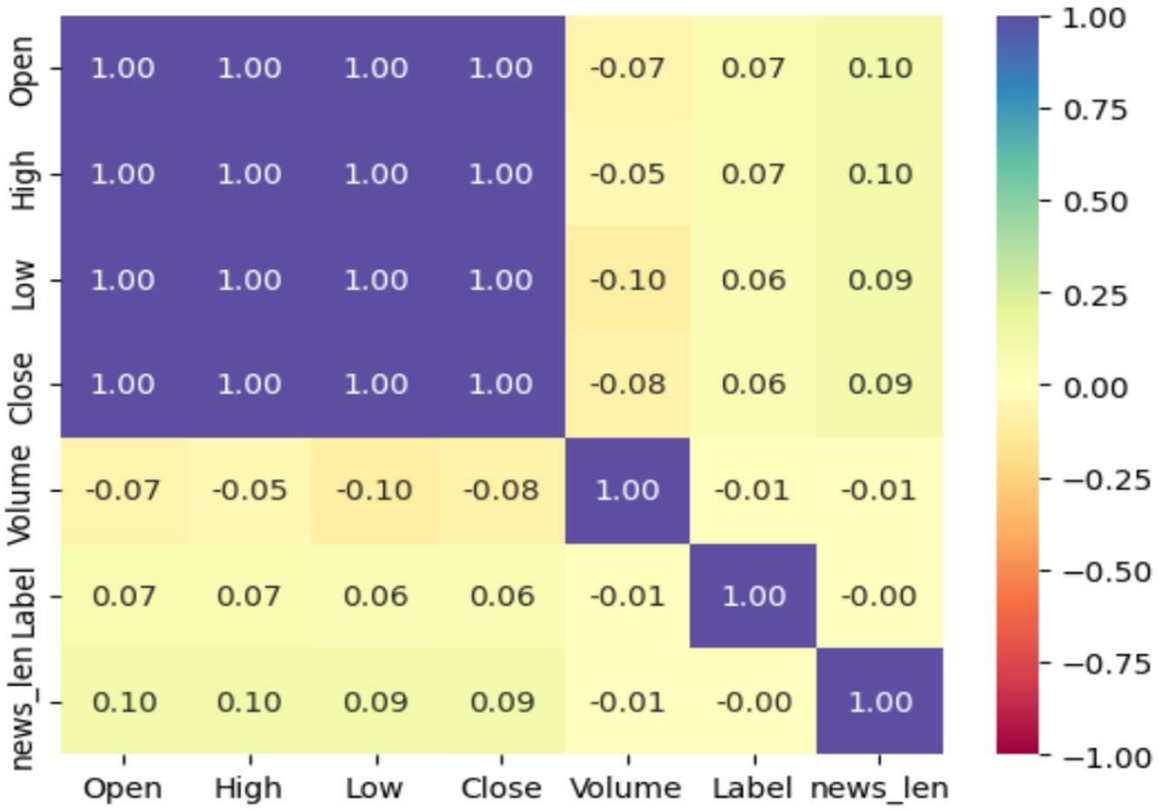
From this , it is clear that the data distribution is not uniform across. Mean is along the 50

```
sns.histplot(data=stock,x="news_len"); #Complete the code to plot a histogram
```



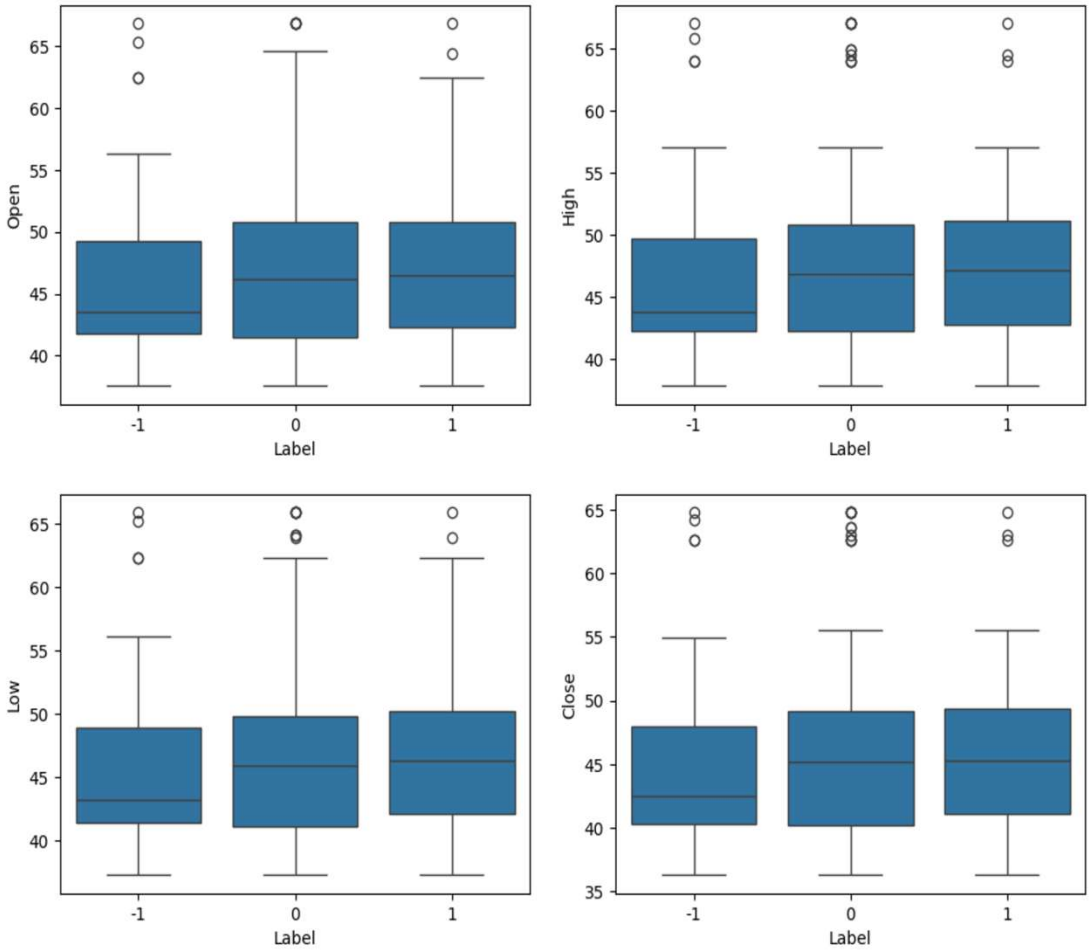
EDA Results

- In the Bi variate analysis



A **heat chart** (more commonly called a **heatmap**) is a type of **data visualization** that shows the **magnitude of values** across a matrix as **colors**. It's perfect for visualizing things like **correlation matrices**, **confusion matrices**, or **feature relationships**.

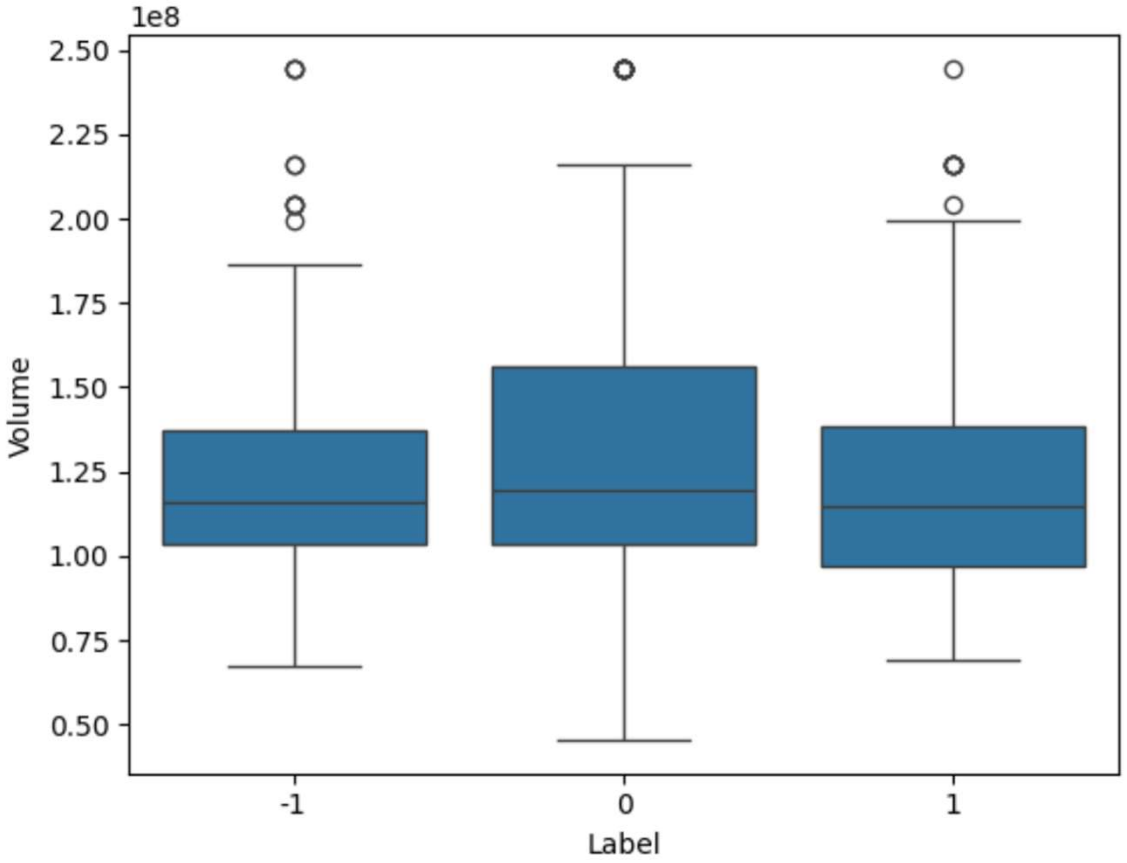
EDA Results



There were many outliers on all the open , High,Low,Close.

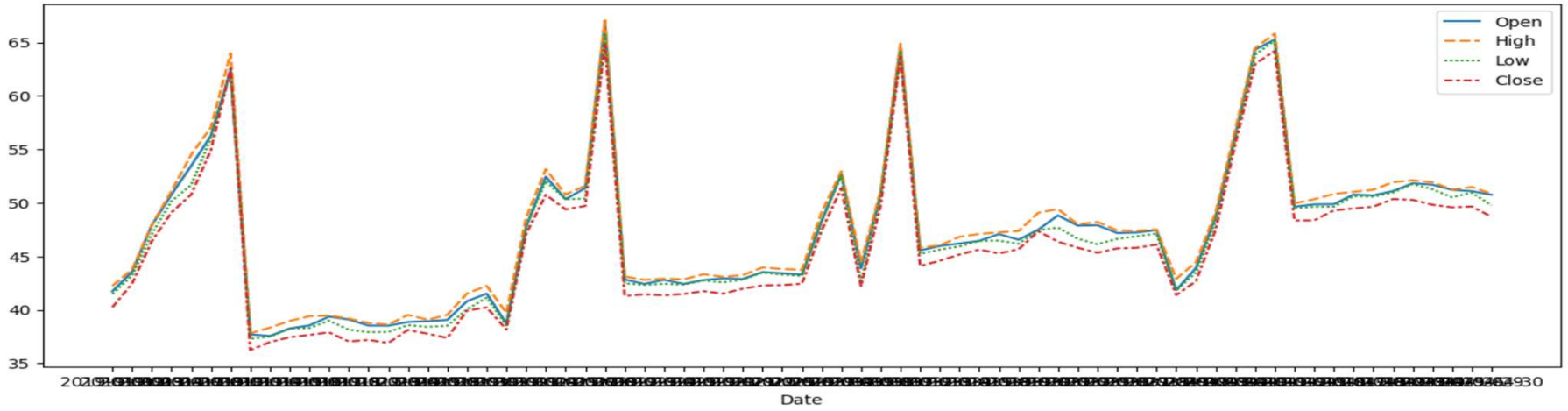
•Each shows how a stock feature (Open, High, Low, Close) varies across different Label categories. Makes it easy to spot **differences in distribution** or **outliers** based on labels (e.g., price going up/down)

EDA Results



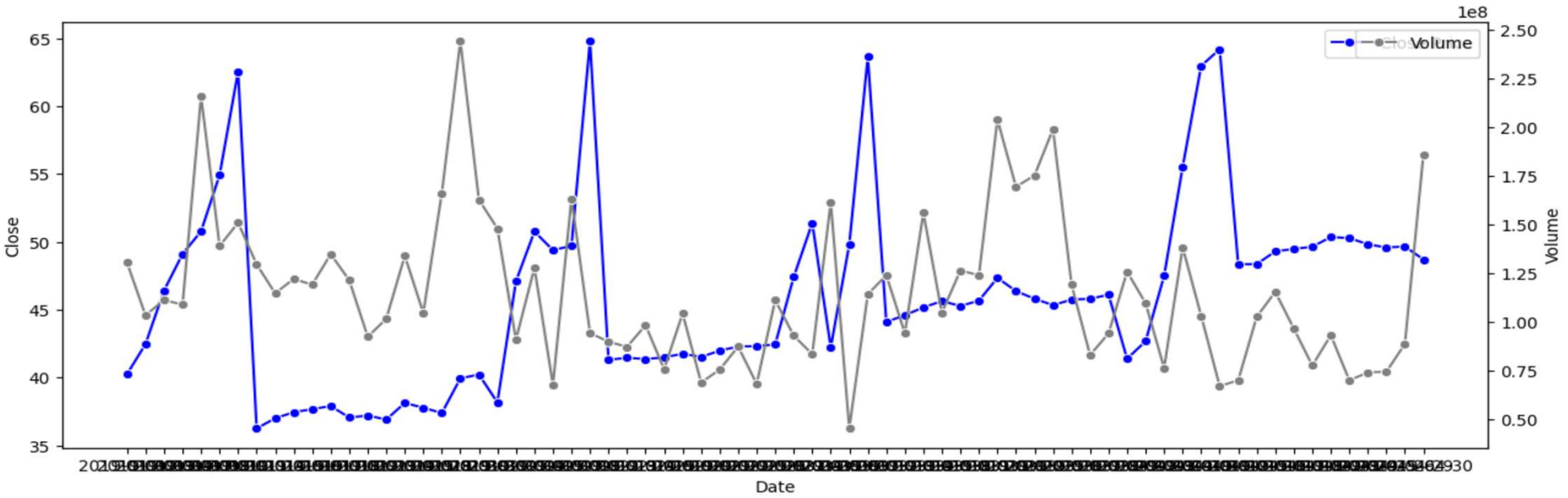
Compares **stock trading volume** across different labels .
Useful to see if certain labels are associated with **higher or more variable volume**.

EDA Results



Date	Open	High	Low	Close	Volume
2019-01-02	41.740002	42.244999	41.482498	40.246914	130672400.0
2019-01-03	43.570000	43.787498	43.222500	42.470604	103544800.0
2019-01-04	47.910000	47.919998	47.095001	46.419842	111448000.0
2019-01-07	50.792500	51.122501	50.162498	49.110790	109012000.0
2019-01-08	53.474998	54.507500	51.685001	50.787209	216071600.0

EDA Results



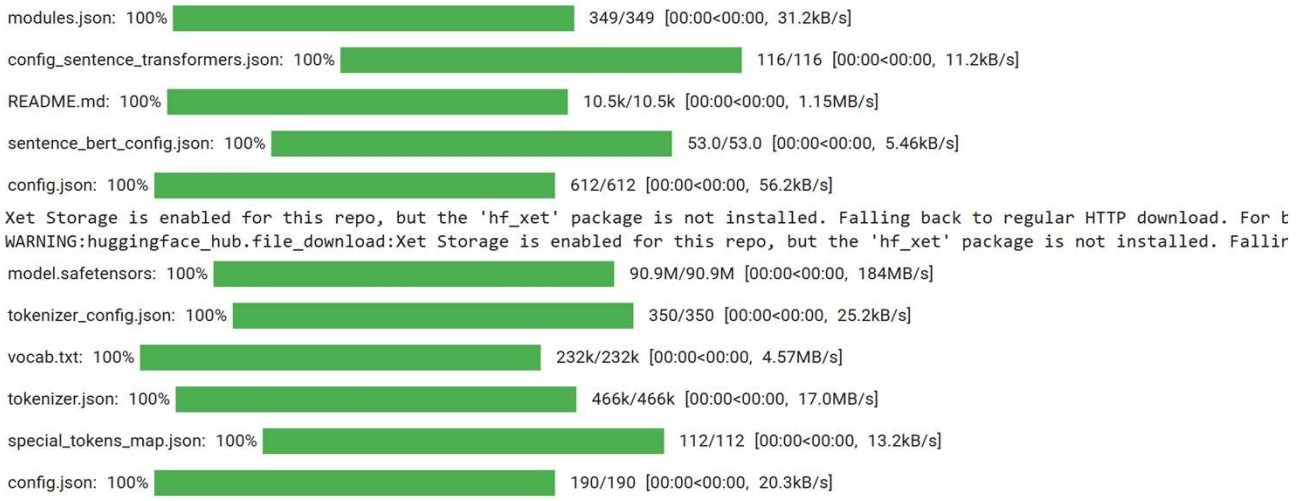
- **Volume Spikes ≠ Price Spikes**: High volume doesn't always mean big price movement — sometimes price remains flat.
- **Sharp price movements** often occur with **volume surges**, suggesting **market reactions** to external factors.
- There's a consistent **range** in volume, but **price shows higher variability**.

EDA Results

	Date
count	349
unique	71
top	2019-01-03
freq	28

```
Train data shape (286, 10)
Validation data shape (21, 10)
Test data shape  (42, 10)
Train label shape (286,)
Validation label shape (21,)
Test label shape  (42,)
```

EDA Results




```
Time taken  31.85196614265442
```

```
print(X_train_gl.shape, X_val_gl.sl
 #Complete the code to print the sl
```

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

```
Time taken  0.5910465717315674
```

```
[ ] print(X_train_wv.shape, X_val_wv.shape, X_test_wv.shape)
```

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

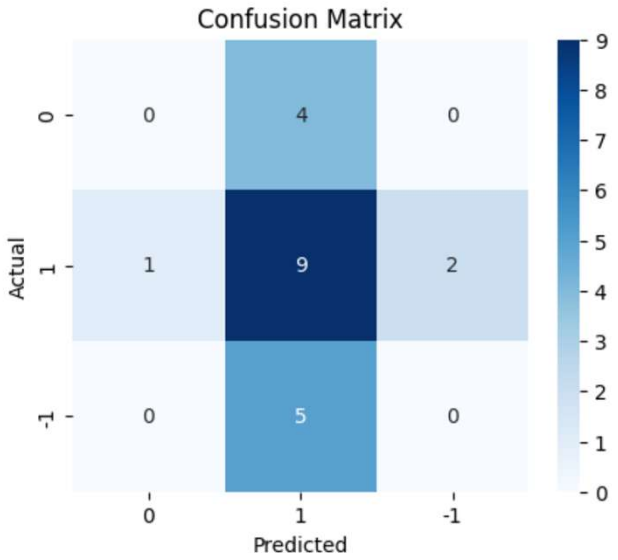
Data Preprocessing

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

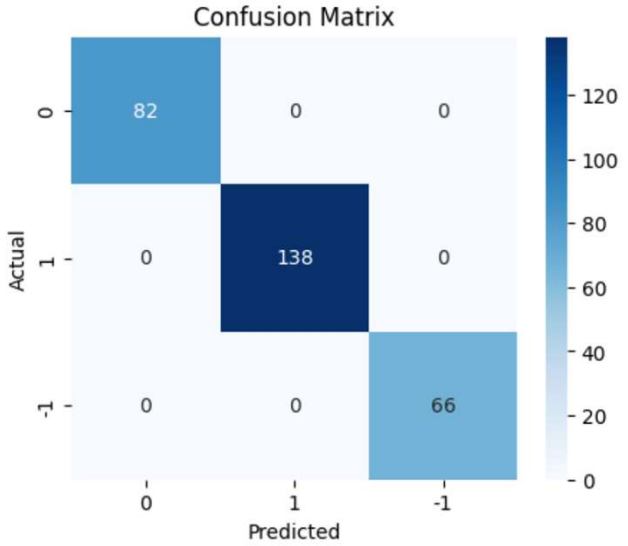
Each news content has been converted to a 384-dimensional vector.

Sentiment Analysis - Model Evaluation Criterion

Base Model - **Word2Vec**



Word2Vec is a popular technique in natural language processing (NLP) for **transforming words into vector representations**— essentially, turning words into numbers that a machine can understand, while capturing their meaning and relationships.



Validation performance:

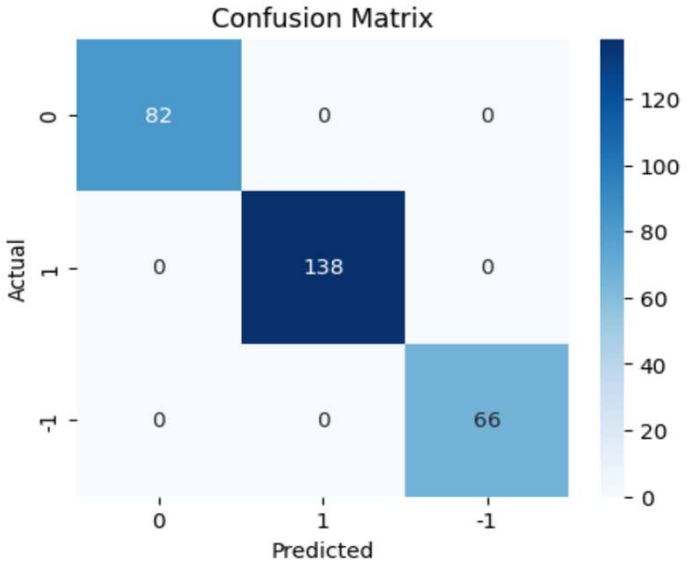
	Accuracy	Recall	Precision	F1
0	0.428571	0.428571	0.285714	0.342857

Training performance:

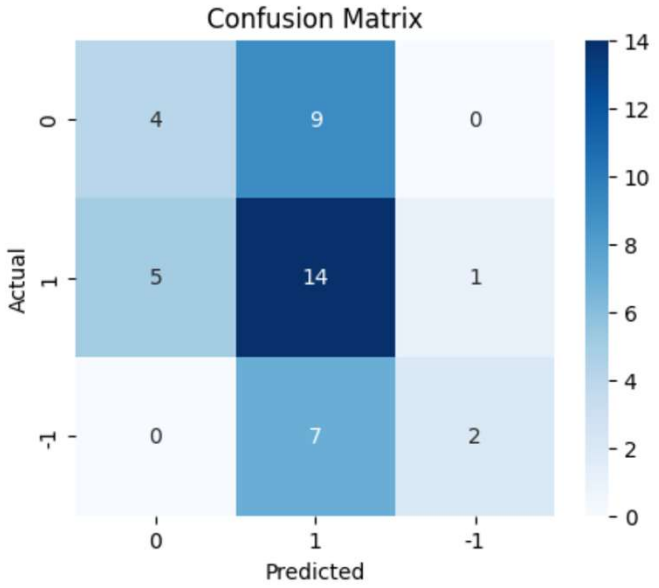
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Sentiment Analysis - Model Evaluation Criterion

Base Model - GloVe



GloVe is an unsupervised learning algorithm developed by researchers at Stanford for **generating word embeddings**.
It combines the **global statistical information** of a corpus (like how often words co-occur) with the **local context** of words to produce word vectors.



Training performance:

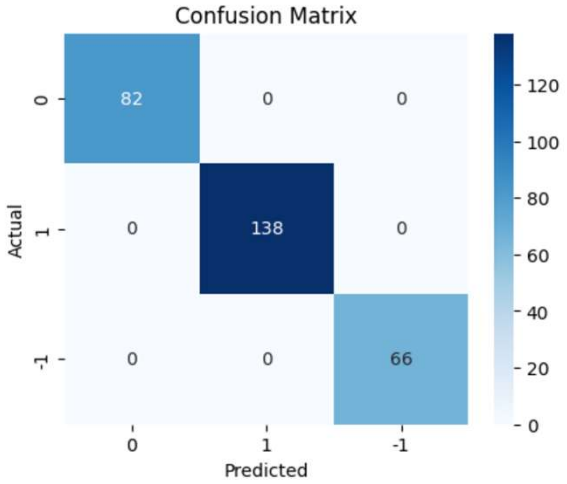
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation performance:

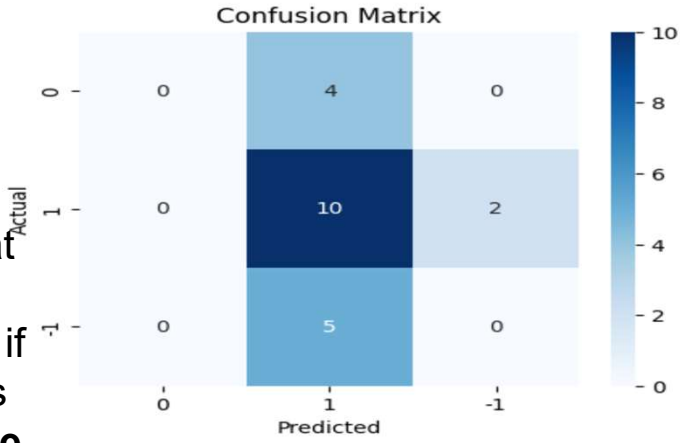
	Accuracy	Recall	Precision	F1
0	0.380952	0.380952	0.326531	0.351648

Sentiment Analysis - Model Evaluation Criterion

Base Model - Sentence Transformer



A **Sentence Transformer** is a type of model that converts **entire sentences** (not just individual words) into **dense vector embeddings** that capture their **semantic meaning**. These embeddings are designed so that **similar sentences are close together in vector space**, even if they use different words. It builds on the **Transformer architecture**, especially BERT and its variants.



Training performance:

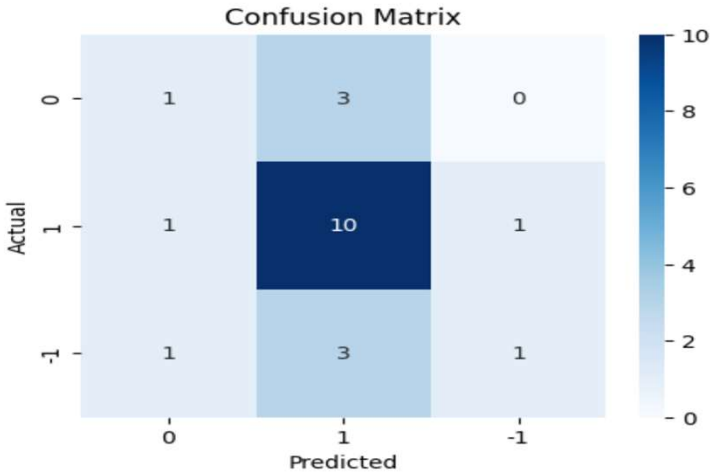
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation performance:

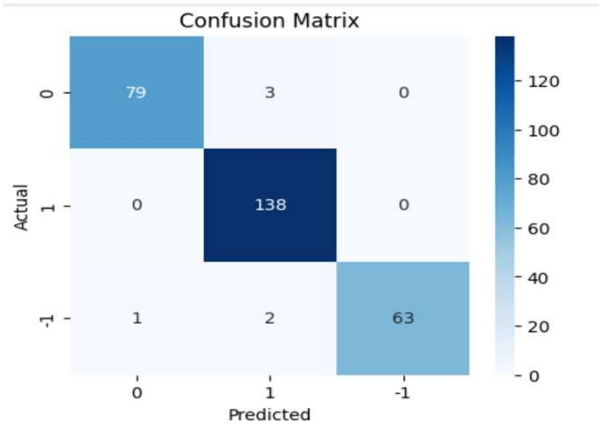
	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.300752	0.368664

Sentiment Analysis - Model Evaluation Criterion

Tuned Model - GloVe



Base glove model is finely tuned for the better. Pre-trained models are more powerful .Tuning the chart makes More accurate.



Validation performance:

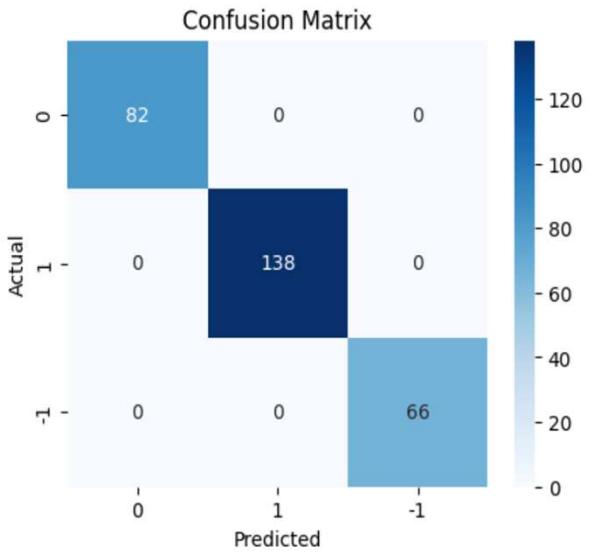
	Accuracy	Recall	Precision	F1
0	0.571429	0.571429	0.539683	0.530612

Training performance:

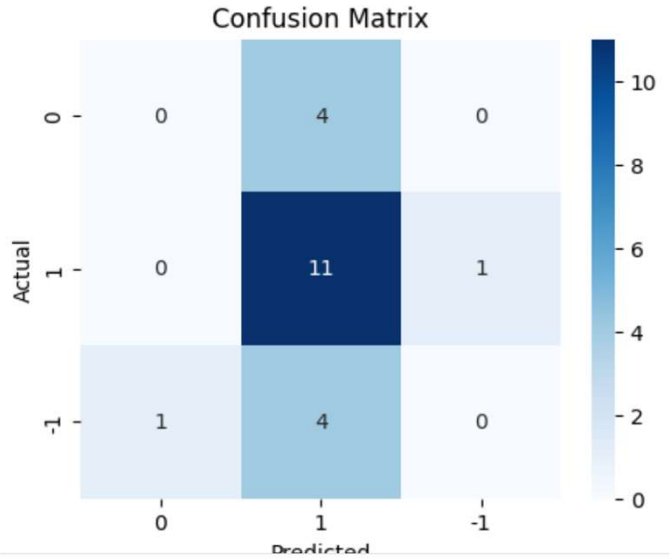
	Accuracy	Recall	Precision	F1
0	0.979021	0.979021	0.979545	0.978968

Sentiment Analysis - Model Evaluation Criterion

Tuned Model - Word2Vec



Word2Vec is static , so even a tuned model cant understand contextual word meanings, tuning requires ample data to be effective.word2Vec works well in resource-limited environments.



Training performance:

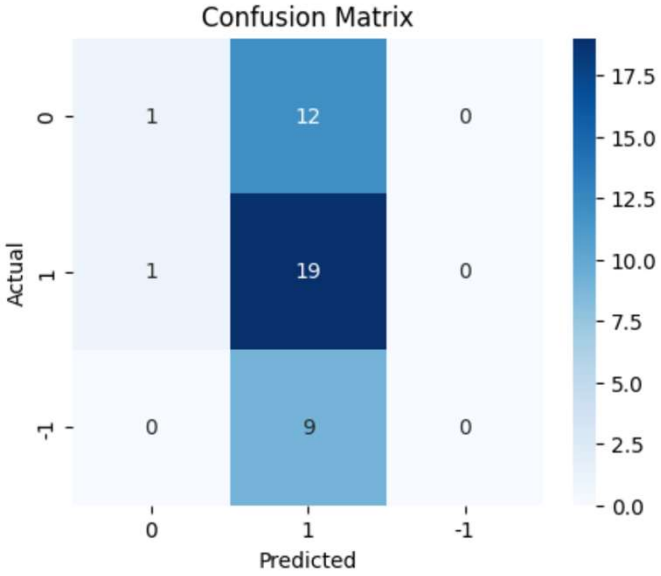
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Validation performance:

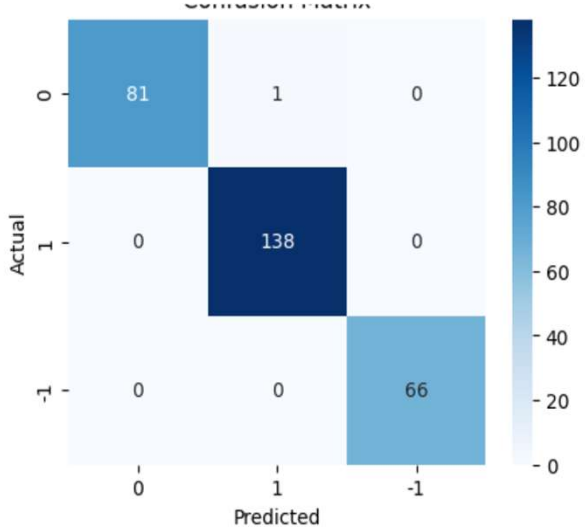
	Accuracy	Recall	Precision	F1
0	0.52381	0.52381	0.330827	0.40553

Sentiment Analysis - Model Evaluation Criterion

Tuned Model - Sentence Transformer



A tuned Sentence Transformer is a pre-trained model
that has been fine tuned on a specific task or domain.
For semantic search , tuned embedding gets better.,Compact models with better accuracy



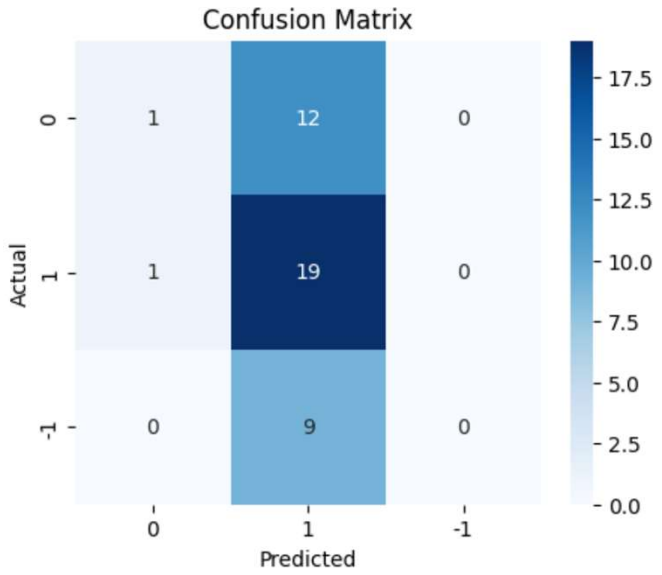
Validation performance:

	Accuracy	Recall	Precision	F1
0	0.571429	0.571429	0.326531	0.415584

Training performance:

	Accuracy	Recall	Precision	F1
0	0.996503	0.996503	0.996529	0.996499

Sentiment Analysis - Model Building



Validation performance:

	Accuracy	Recall	Precision	F1
0	0.571429	0.571429	0.326531	0.415584

Training performance:

	Accuracy	Recall	Precision	F1
0	0.996503	0.996503	0.996529	0.996499

Sentiment Analysis - Model Improvement

comparing the **training performance** of **six different models**, which vary by:
•The type of word embeddings used (Word2Vec, GloVe, Sentence Transformer),
•Whether the model is **base** or **tuned**.

This code lets you **compare model performance side by side**, making it easy to see:
•Which embeddings perform better,
•Whether tuning helps,
•Which model is overall best for your training set.

•**B**ase_train_wv, base_train_gl, base_train_st, etc. are likely **DataFrames or Series** that store training metrics like accuracy, precision, recall, F1-score, etc.
•.T means **transpose**, switching rows and columns.
•Why? Because each model's metrics are probably **in a column**, and to line them up side by side, we need to **transpose** so the metrics become rows.
•pd.concat([...], axis=1):
•Combines all these transposed models **column-wise** into one big DataFrame.
•This way, you end up with **a single DataFrame** where:
    •Rows = metrics (e.g., accuracy, precision, etc.)
    •Columns = different models

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	0.979021	0.996503
Recall	1.0	1.0	1.0	1.0	0.979021	0.996503
Precision	1.0	1.0	1.0	1.0	0.979545	0.996529
F1	1.0	1.0	1.0	1.0	0.978968	0.996499

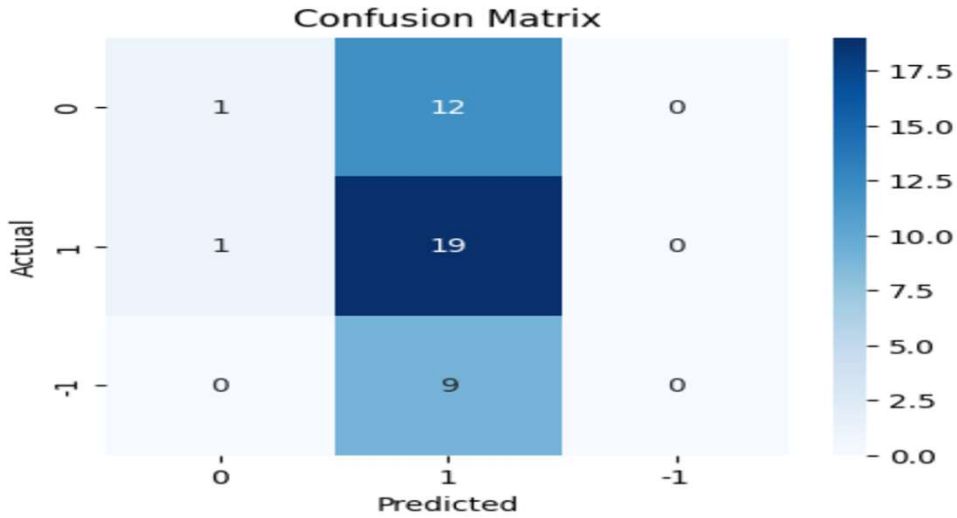
Sentiment Analysis – Final Model

Validation performance comparison:

	Base Model (Word2Vec)	Base Model (GloVe)	Base Model (Sentence Transformer)	Tuned Model (Word2Vec)	Tuned Model (GloVe)	Tuned Model (Sentence Transformer)
Accuracy	0.428571	0.380952	0.476190	0.523810	0.571429	0.571429
Recall	0.428571	0.380952	0.476190	0.523810	0.571429	0.571429
Precision	0.285714	0.326531	0.300752	0.330827	0.539683	0.326531
F1	0.342857	0.351648	0.368664	0.405530	0.530612	0.415584

Test performance for the final model

	Accuracy	Recall	Precision	F1
0	0.47619	0.47619	0.380952	0.342857



Weekly News Summarization

```
weekly_grouped
```

	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...
6	2019-02-17	This week, the European Union's second highes...
7	2019-02-24	This news article discusses progress towards ...
8	2019-03-03	The Dow Jones Industrial Average and other ma...
9	2019-03-10	Spotify, the world's largest paid music strea...
10	2019-03-17	The United States opposes France's digital se...
11	2019-03-24	Facebook's stock price dropped more than 3% o...
12	2019-03-31	This news article reports that the S&P 500 In...
13	2019-04-07	Apple and other consumer brands, including LV...
14	2019-04-14	In March, mobile phone shipments to China dro...
15	2019-04-21	The chairman of Taiwan's Foxconn, Terry Gou, ...
16	2019-04-28	Taiwan's export orders continued to decline f...
17	2019-05-05	Spotify reported better-than-expected Q1 reve...

	Key Events
0	1. Apple's Q1 revenue warning led to significa...
1	1. Sprint and Samsung plan to release 5G smart...
2	This news article covers various events relat...
3	Title: Global Markets and Business News: Swis...
4	1. Caterpillar Inc reports lower-than-expected...

	Date	News	Key Events	model_response_parsed
0	2019-01-06	The tech sector experienced a significant dec...	1. Apple's Q1 revenue warning led to significa...	{}
1	2019-01-13	Sprint and Samsung plan to release 5G smartph...	1. Sprint and Samsung plan to release 5G smart...	{}
2	2019-01-20	The U.S. stock market declined on Monday as c...	This news article covers various events relat...	{}
3	2019-01-27	The Swiss National Bank (SNB) governor, Andre...	Title: Global Markets and Business News: Swis...	{}
4	2019-02-03	Caterpillar Inc reported lower-than-expected ...	1. Caterpillar Inc reports lower-than-expected...	{}

Content Summarization – Modeling Approach

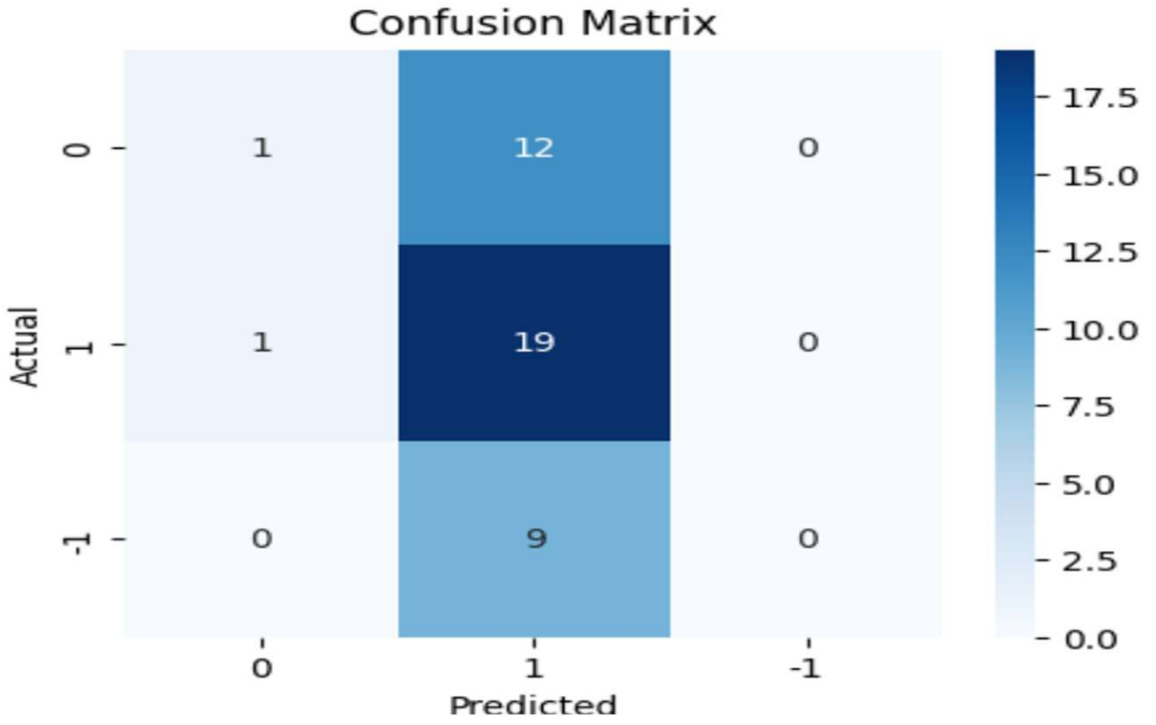
```
model_response_parsed.head()
```

```
print(final_output.shape)
```

```
(18, 2)
```

```
print(data_1.shape)
print(model_response_
```

```
(18, 4)
(18, 0)
```



	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 ...

Inference: Transformer model is best. With the transformer, the task is achieved.

APPENDIX

Data Background and Contents

```
stock.head() # Complete the code to check the first 5 rows of the data
```

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 349 entries, 0 to 348
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Date    349 non-null    object 
 1   News    349 non-null    object 
 2   Open    349 non-null    float64
 3   High    349 non-null    float64
 4   Low     349 non-null    float64
 5   Close   349 non-null    float64
 6   Volume  349 non-null    int64  
 7   Label   349 non-null    int64  
dtypes: float64(4), int64(2), object(2)
memory usage: 21.9+ KB
```

	Date	Open	High	Low	Close	Volume	Label
count	349	349.000000	349.000000	349.000000	349.000000	3.490000e+02	349.000000
mean	2019-02-16 16:05:30.085959936	46.229233	46.700458	45.745394	44.926317	1.289482e+08	-0.054441
min	2019-01-02 00:00:00	37.567501	37.817501	37.305000	36.254131	4.544800e+07	-1.000000
25%	2019-01-14 00:00:00	41.740002	42.244999	41.482498	40.246914	1.032720e+08	-1.000000
50%	2019-02-05 00:00:00	45.974998	46.025002	45.639999	44.596924	1.156272e+08	0.000000
75%	2019-03-22 00:00:00	50.707500	50.849998	49.777500	49.110790	1.511252e+08	0.000000
max	2019-04-30 00:00:00	66.817497	67.062500	65.862503	64.805229	2.444392e+08	1.000000
std	NaN	6.442817	6.507321	6.391976	6.398338	4.317031e+07	0.715119

Great Learning

Happy Learning !