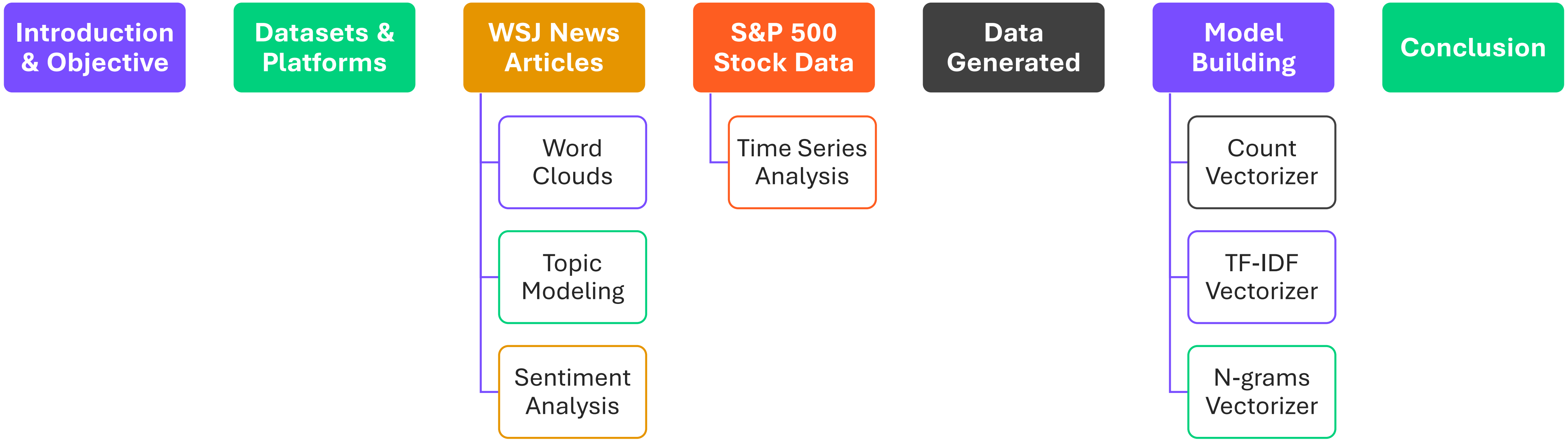
BUAN 6342 NLP Project

Predicting the Changes of S&P 500 Stock Price based on WSJ News Articles – NLP & ML Approaches

Presenter: Min Shi

Date: December 1st, 2022

Overview



Introduction & Objective

Datasets & Platforms

WSJ News Articles
- Word Clouds
- Topic Modeling
- Sentiment Analysis

S&P 500 Stock Data
- Time Series Analysis

Data Generated

Model Building
- Count Vectorizer
- TF-IDF Vectorizer
- N-grams Vectorizer

Conclusion

INTRODUCTION

Newspapers: disseminating information to a broad audience, affect audience's actions.

- 58% of adults aged 18-34 and 60% aged over 35 read a newspaper
- The Wall Street Journal (WSJ): an American business-focused, international daily newspaper

Stock index: reflecting market and companies' operations status are.

- The Standard and Poor's 500 (S&P 500): a stock market index tracking the stock performances of 500 large companies listed on the stock exchanges in the U.S. and is one of the most followed equity indices

OBJECTIVES

The U.S. is the world's largest exporter and importer of goods and services. Trade is critical to U.S. companies.

In this project, I investigate the WSJ news articles related to "U.S. trade" and use the matrix of token counts based on the collection of news article text to predict how the U.S. stock market and U.S. companies operate measured by S&P 500 stock index.

DATASETS

The News Articles:

- Scraped from the Wall Street Journal (WSJ)
- Period: January 1, 2018 - October 31, 2022
- Python is used to scrape newspaper articles

The S&P 500 Stock Index Data:

- Scraped from the yahoo finance website
- Period: January 1, 2018 - October 31, 2022
- R package quantmod is utilized

PLATFORMS

R:

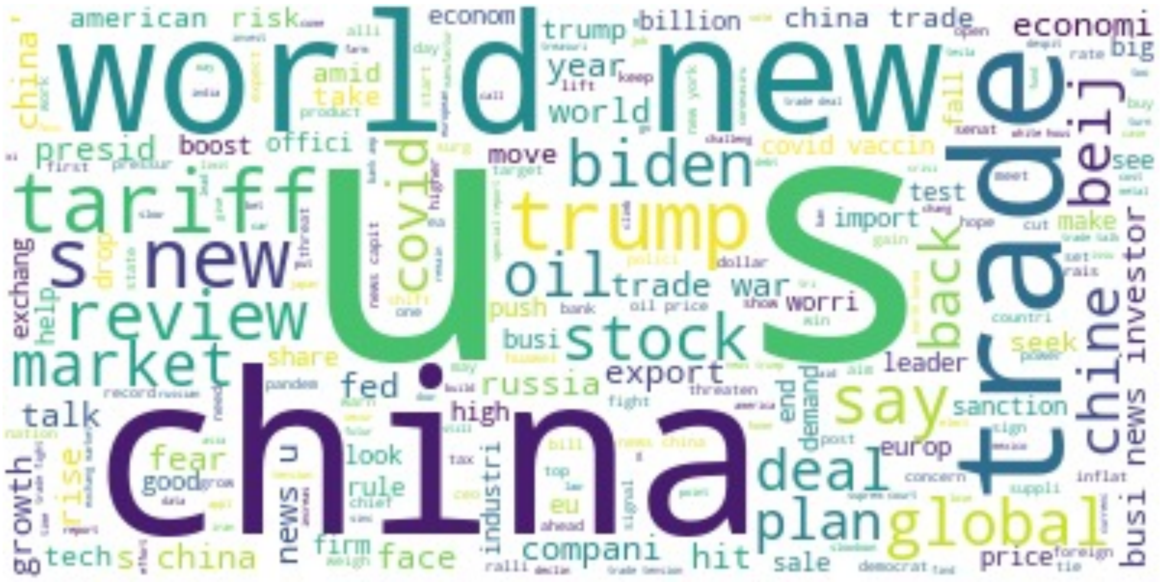
- Utilized to scrape S&P 500 stock index data

Python:

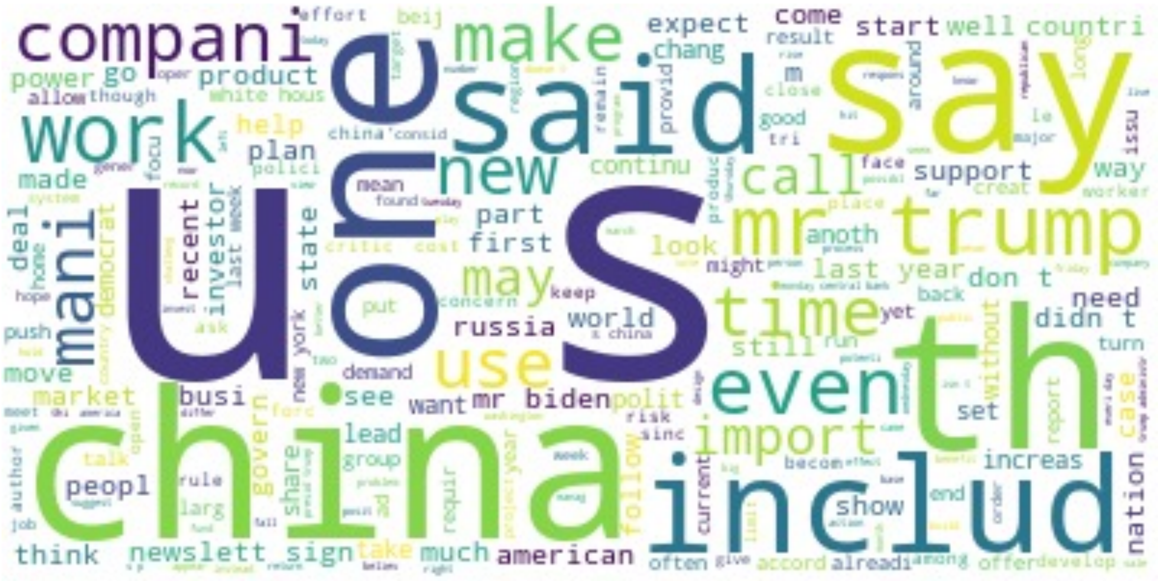
- Used to scrape newspaper articles from the Wall Street Journal website (Time consuming, cost several days)
- The main platform for data analytics, using the following packages:
  - Data Processing: Numpy, pandas, os, sys, cvs, string, pickle
  - Data Visualization: matplotlib, seaborn
  - Topic Modeling: wordcloud, sklearn, genism
  - Sentiment Analysis: nltk, vaderSentiment, textblob
  - Time Series Data Analysis: sys, statsmodels, scipy
  - NLP: nltk, statsmodels, sklearn, spicy, etc.

WSJ News Articles – Word Clouds

Newspaper Titles



Newspaper Texts



Frequent words in Newspaper Titles:

- U.S., China, trade, world, new, tariff, trump, biden, stock, market, global, Beijing and Chinese

Frequent words in Newspaper Text:

- U.S., China, one, say, said, include, trump, company, work, Trump

WSJ News Articles – Topic Modeling (News Title)

	Topic 1 Market	Topic 2 Global Trade	Topic 3 Economy	Topic 4 Trade & Deal	Topic 5 Covid & Vaccine	Topic 6 Products	Topic 7 Invest	Topic 8 Russia & Ukraine	Topic 9 Trade War	Topic 10 China
0	market	news	hous	trade	covid	new	price	review	trump	china
1	exchang	world	ahead	china	say	bank	oil	biden	tariff	chine
2	stock	china	economi	hit	vaccin	york	stock	russia	trade	news
3	amp	trade	white	talk	state	product	rise	court	china	busi
4	fall	growth	cut	deal	test	china	investor	rule	war	firm
5	china	busi	tax	report	elect	billion	fed	ukrain	steel	maker
6	week	beij	crisi	record	case	tesla	year	plan	presid	north
7	day	capit	set	tension	ceo	dollar	worri	pandem	import	korea
8	data	economi	oil	econom	biden	shift	global	foreign	tariffs	sanction
9	markets	global	say	year	trump	job	fear	work	good	sale

WSJ News Articles – Topic Modeling (News Text)

	Topic 1 Tech Company	Topic 2 Global Trade	Topic 3 International Relations	Topic 4 Politics & Election	Topic 5 School & College	Topic 6 Stock & Inflation	Topic 7 Covid & Vaccine	Topic 8 Russia & Ukraine	Topic 9 Manufacture	Topic 10 Life & Family
0	company	eu	beij	democrat	court	stock	vaccin	manufactur	russia	life
1	tech	north	huawei	republican	school	fed	covid	car	oil	book
2	appl	steel	xi	vote	justic	dollar	health	maker	russian	says
3	employe	mexico	beijing	elect	student	inflat	test	sale	energi	live
4	job	negoti	iran	tax	rule	index	dr	plant	ukrain	old
5	app	agreement	india	senat	judg	rose	viru	electr	ga	play
6	pay	korea	hong	congress	board	fell	drug	auto	ukraine	feel
7	worker	canada	militari	parti	investig	quarter	dose	vehicl	putin	art
8	amazon	tariffs	kong	sen	legal	economist	hospit	factori	ukrainian	famili
9	servic	impos	taiwan	voter	depart	central	coronaviru	ship	sanction	thought

WSJ News Articles – Topic Modeling

10 topics for news titles: *Market, Global Trade, Economy, Trade & Deal, Covid & Vaccine, Products, Invest, Russia & Ukraine, Trade War, China*.

10 topics of news text: *Tech Company, Global Trade, International Relations, Politics & Election, School & College, Stock & Inflation, Covid & Vaccine, Russia & Ukraine, Manufacture, and Life & Family.*

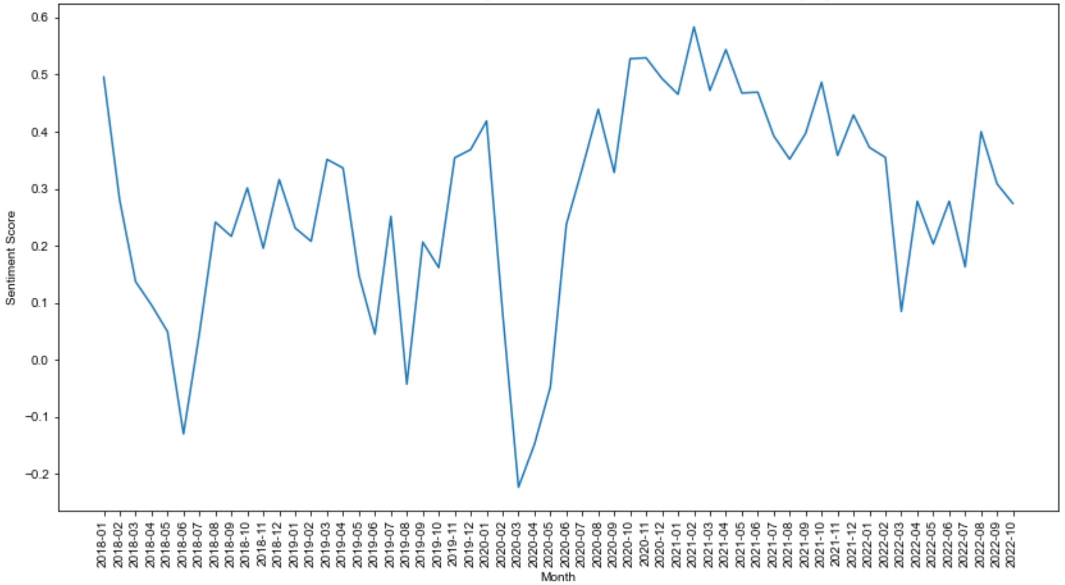
Multiple types of topics covered: economic, market, political, international relations to school and life, etc.

WSJ News Articles – Sentiment Analysis

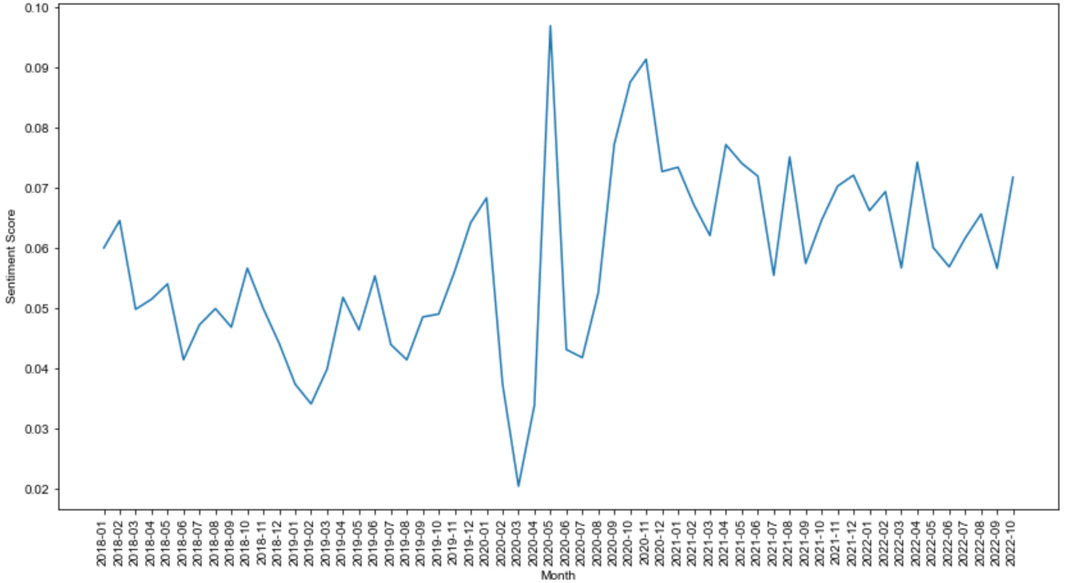
Two Lexicons:

- **VADER (Valence Aware Dictionary for Sentiment Reasoning)**
  - VADER is a dictionary of 4 keys neg, neu, pos and compound
  - Compound corresponds to the sum of the valence score of each word in the lexicon
  - Range of Compound: [-1.0, 1.0]
- **TextBlob:**
  - TextBlob’s output for a polarity task is a float within the range [-1.0, 1.0]
  - Subjectivity/objectivity identification task reports a float within the range [0.0, 1.0]

WSJ News Articles – Sentiment Analysis

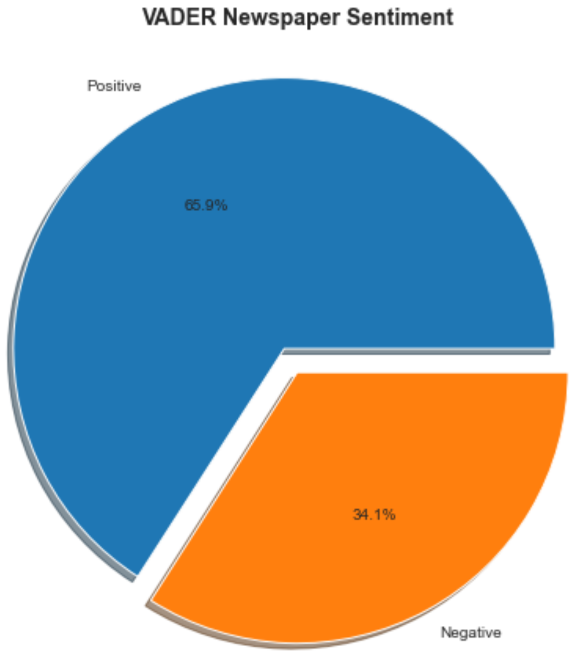


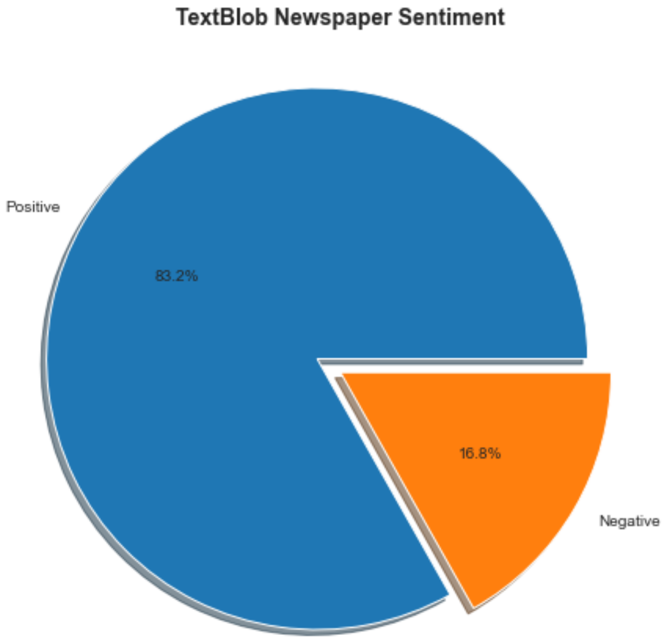
VADER Sentiment Analysis of Wall Street Journal Newspaper



TextBlob Sentiment Analysis of Wall Street Journal Newspaper

WSJ News Articles – Sentiment Analysis





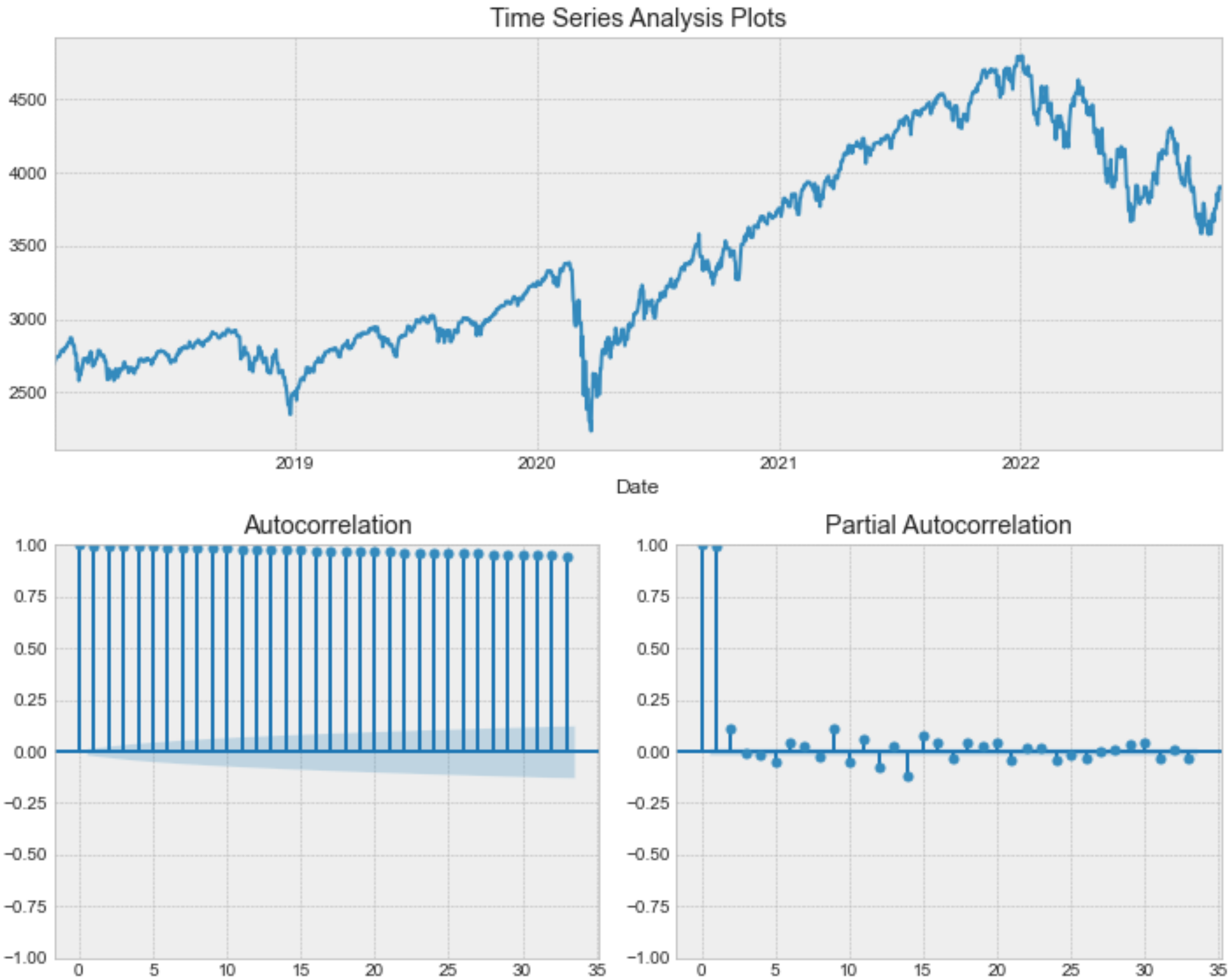
Stock Data Analysis

S&P 500 Stock Index Price

AD Fuller tests to detect the stationarity:

The p-value for the ADF test on S&P500 adjusted stock price is 0.6989.

→ Non-stationary



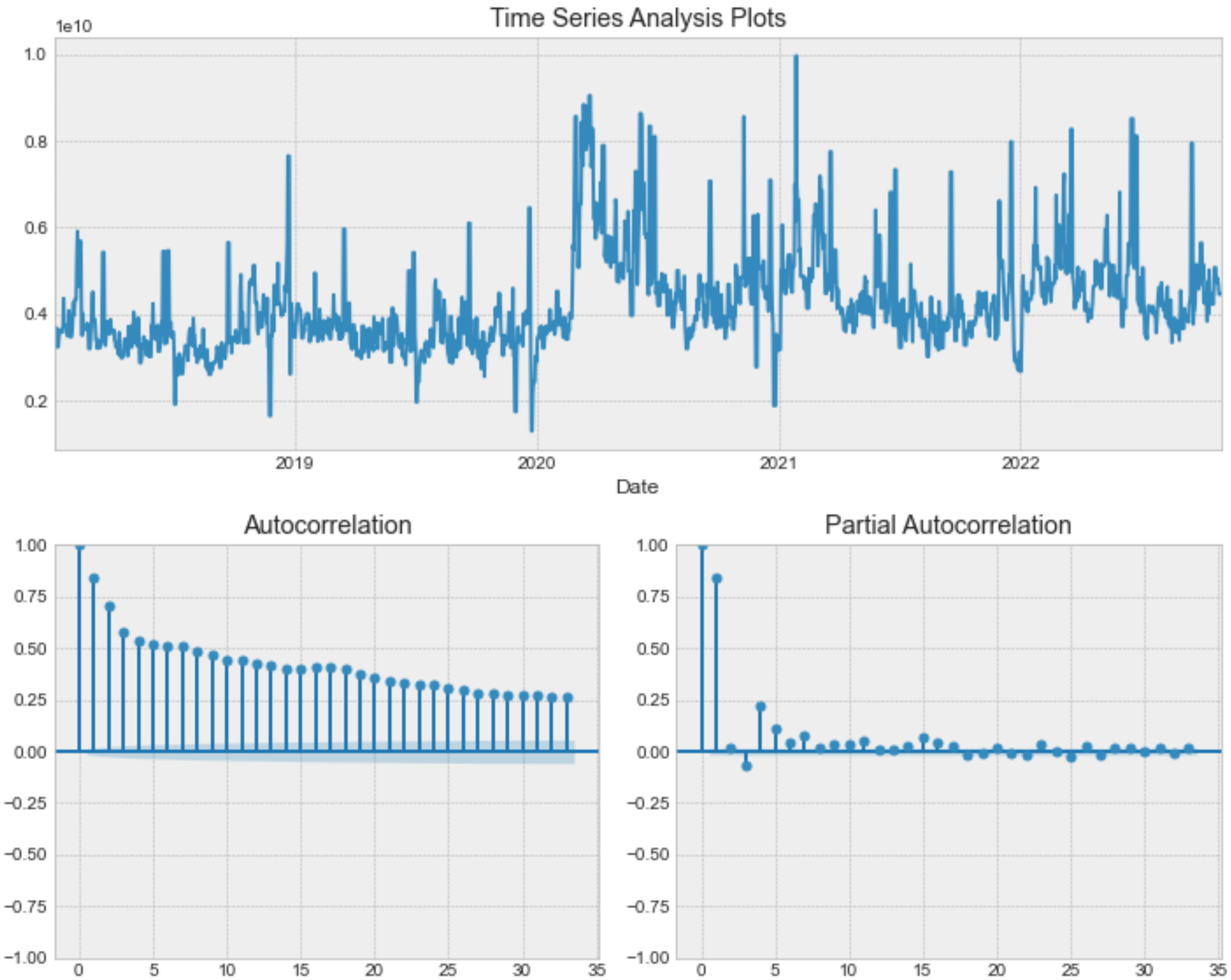
Stock Data Analysis

S&P 500 Total Volume

AD Fuller tests to detect the stationarity:

The p-value for the ADF test on S&P500 total volume is 0.0001.

→ Stationary



Stock Data Analysis

S&P 500 500

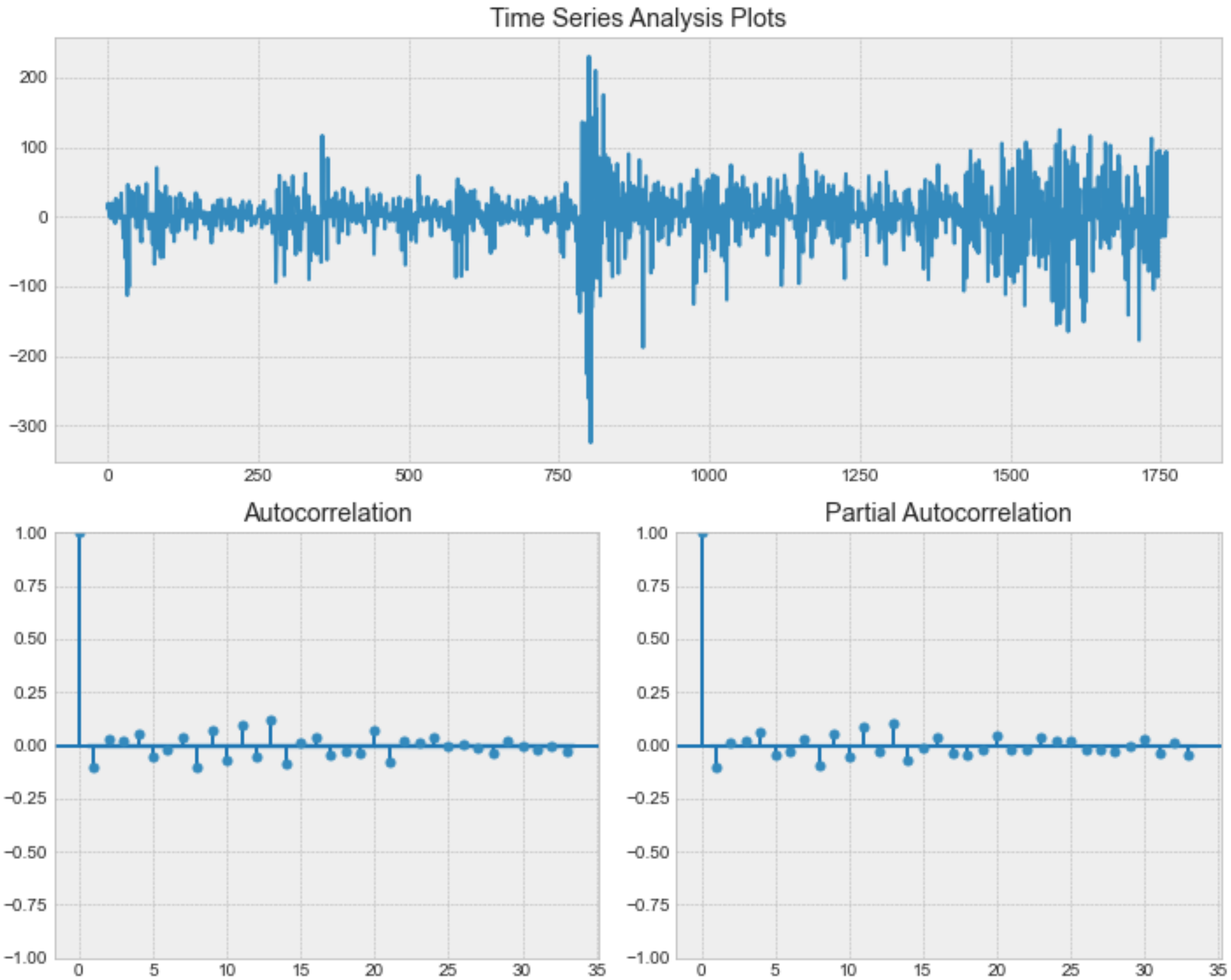
Stock Index Price

1st Difference

AD Fuller tests to detect the stationarity:

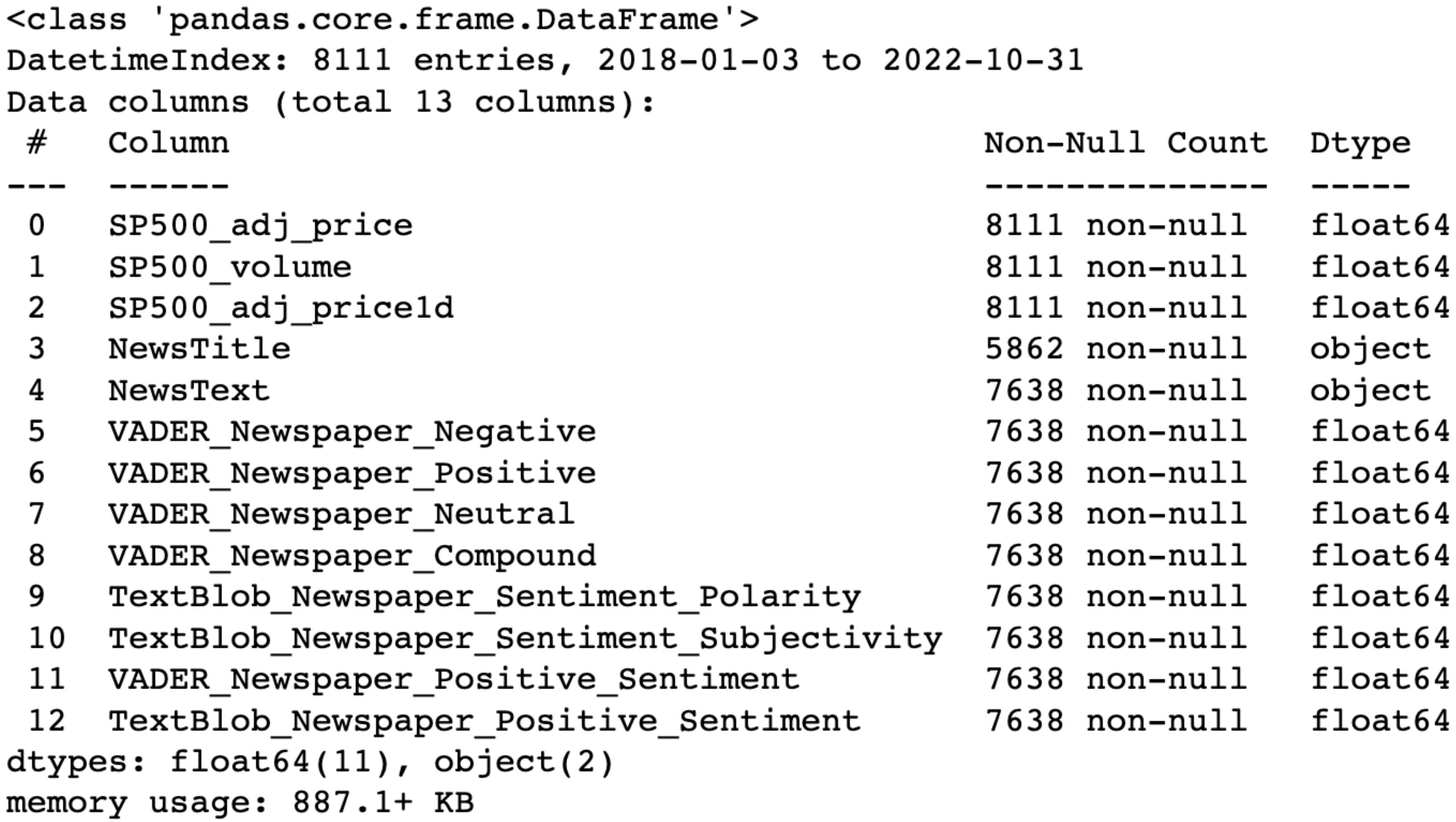
The p-value for the ADF test on S&P500 1st difference is 0.0000.

→ Stationary

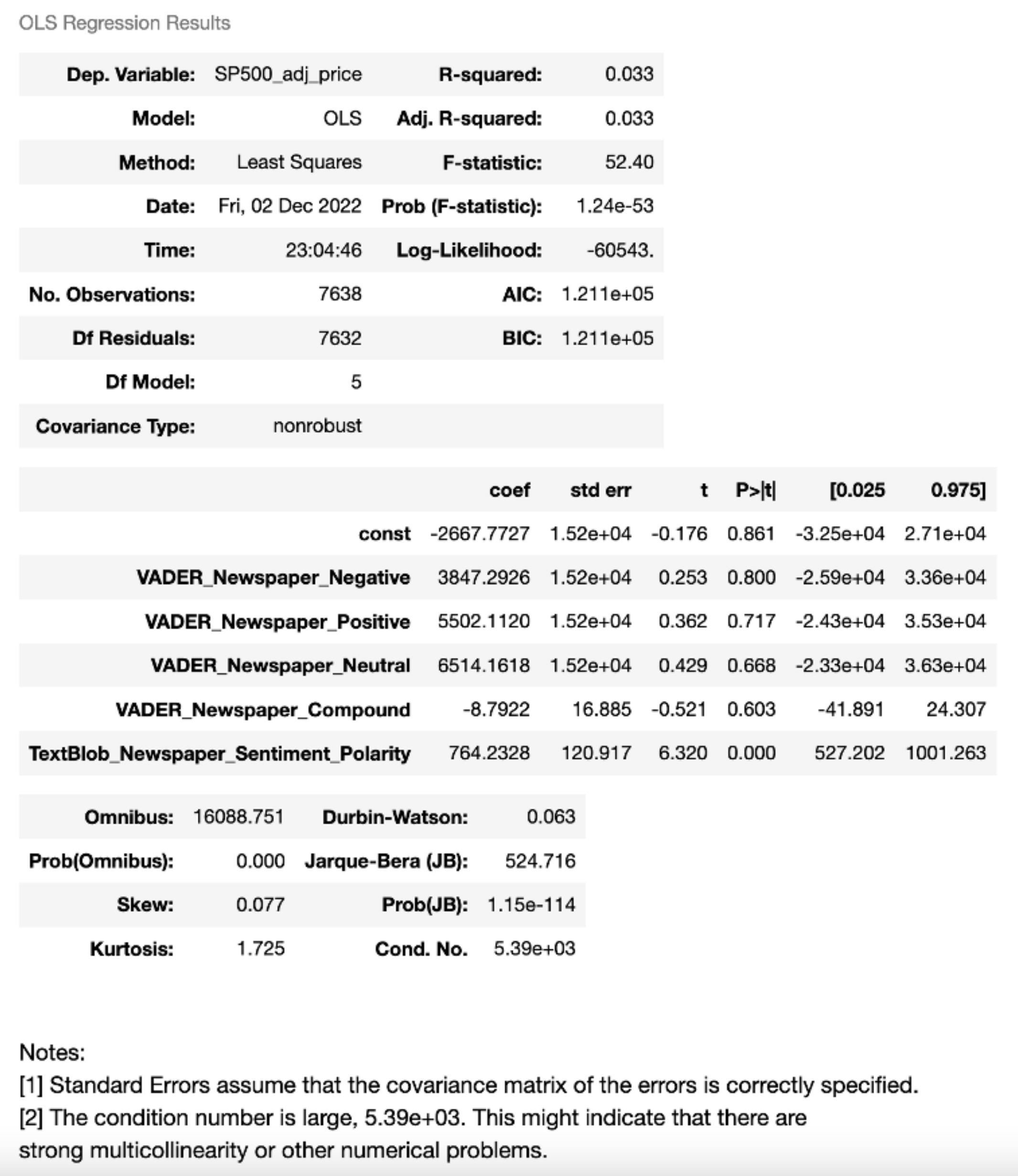


Data Generation

The descriptive statistics of the generated dataset:



```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 8111 entries, 2018-01-03 to 2022-10-31
Data columns (total 13 columns):
 #   Column                                  Non-Null Count  Dtype  
---  ------                                  --------------  -----  
 0   SP500_adj_price                         8111 non-null   float64
 1   SP500_volume                            8111 non-null   float64
 2   SP500_adj_price1d                       8111 non-null   float64
 3   NewsTitle                               5862 non-null   object 
 4   NewsText                                7638 non-null   object 
 5   VADER_Newspaper_Negative                7638 non-null   float64
 6   VADER_Newspaper_Positive                7638 non-null   float64
 7   VADER_Newspaper_Neutral                 7638 non-null   float64
 8   VADER_Newspaper_Compound                7638 non-null   float64
 9   TextBlob_Newspaper_Sentiment_Polarity   7638 non-null   float64
 10  TextBlob_Newspaper_Sentiment_Subjectivity  7638 non-null   float64
 11  VADER_Newspaper_Positive_Sentiment      7638 non-null   float64
 12  TextBlob_Newspaper_Positive_Sentiment   7638 non-null   float64
dtypes: float64(11), object(2)
memory usage: 887.1+ KB
```



OLS Regression Results

Dep. Variable:	SP500_adj_price	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.033
Method:	Least Squares	**F-statistic:**	52.40
Date:	Fri, 02 Dec 2022	**Prob (F-statistic):**	1.24e-53
Time:	23:04:46	**Log-Likelihood:**	-60543.
No. Observations:	7638	**AIC:**	1.211e+05
Df Residuals:	7632	**BIC:**	1.211e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2667.7727	1.52e+04	-0.176	0.861	-3.25e+04	2.71e+04
VADER_Newspaper_Negative	3847.2926	1.52e+04	0.253	0.800	-2.59e+04	3.36e+04
VADER_Newspaper_Positive	5502.1120	1.52e+04	0.362	0.717	-2.43e+04	3.53e+04
VADER_Newspaper_Neutral	6514.1618	1.52e+04	0.429	0.668	-2.33e+04	3.63e+04
VADER_Newspaper_Compound	-8.7922	16.885	-0.521	0.603	-41.891	24.307
TextBlob_Newspaper_Sentiment_Polarity	764.2328	120.917	6.320	0.000	527.202	1001.263

Omnibus:	16088.751	**Durbin-Watson:**	0.063
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	524.716
Skew:	0.077	**Prob(JB):**	1.15e-114
Kurtosis:	1.725	**Cond. No.**	5.39e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.39e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Model Building

Exploring Stock Price vs. Newspaper Sentiment

Linear Regression Model:
SP500_adj_price =
VADER_Newspaper_Negative +
VADER_Newspaper_Positive +
VADER_Newspaper_Neutral +
VADER_Newspaper_Compound +
TextBlob_Newspaper_Sentiment_Polarity

Model Building

Predicting Stock Price Change Using Newspaper Text:

Dependent Variable:

SP500_adj_price_increase with 1 representing increase and 0 representing decrease in S&P 500 stock price

Independent Variables:

A matrix of token counts generated by CountVectorizer, TfidfVectorizer and n-grams CountVectorizer from the collection of news article content.

Model Building – Count Vectorizer

Model	ROC_AUC	Training Score	Testing Score
Logistic Regression	0.5552	0.9998	0.5796
K-Nearest Neighbors (KNN)	0.5167	0.7018	0.5634
Decision Tree	0.5364	0.9998	0.5586
Random Forest	0.5246	0.9998	0.6099
Stochastic Gradient Descent (SGD)	0.5487	0.9921	0.5681
Naive Bayes	0.5750	0.7758	0.5780
Support Vector Machine (SVM)	0.5502	0.9997	0.5702

Through ROC AUC value, the Naïve Bayes model performs best, with 0.5750 ROC AUC value.

Based on testing accuracy, random forest performs best, with 0.6099 accuracy.

Model Building – TF-IDF Vectorizer

Model	ROC_AUC	Training Score	Testing Score
Logistic Regression	0.5429	0.7703	0.6105
K-Nearest Neighbors (KNN)	0.5540	0.7280	0.5906
Decision Tree	0.5157	0.9998	0.5429
Random Forest	0.5407	0.9998	0.6084
Stochastic Gradient Descent (SGD)	0.5643	0.8986	0.6068
Naive Bayes	0.5044	0.6231	0.6042
Support Vector Machine (SVM)	0.5535	0.8355	0.6126

Through ROC AUC value, the Stochastic Gradient Descent (SGD) model performs best, with 0.5643 ROC AUC value.

Based on testing accuracy, Support Vector Machine (SVM) performs best, with 0.6126 accuracy.

Model Building – N-Grams Vectorizer

Model	ROC_AUC	Training Score	Testing Score
Logistic Regression	0.5632	0.9998	0.5890
K-Nearest Neighbors (KNN)	0.5078	0.6985	0.5623
Decision Tree	0.5469	0.9998	0.5702
Random Forest	0.5233	0.9998	0.6079
Stochastic Gradient Descent (SGD)	0.5597	0.9749	0.5853
Naive Bayes	0.5822	0.7987	0.5895
Support Vector Machine (SVM)	0.5620	0.9997	0.5848

Through ROC AUC value, the Naive Bayes model performs best, with 0.5822 ROC AUC value.

Based on testing accuracy, Random Forest performs best, with 0.6079 accuracy.

CONCLUSIONS

- The content of newspaper articles could be used to predict the changes of stock price index.

- Naïve Bayes model is the best model according to ROC AUC value and Random Forest model has the best performance according to testing score.

- There is no big difference on the model accuracy utilizing different vectorizers to generate the matrix of token counts.

References

Calderon, Pio. (2017, April 10). VADER Sentiment Analysis Explained. https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9#:~:text=VADER%20(Valence%20Aware%20Dictionary%20for,intensity%20(strength)%20of%20emotion (Accessed November 30, 2022).

Keita, Zoumana. (n.d.) Social Media Sentiment Analysis In Python With VADER — No Training Required! https://towardsdatascience.com/social-media-sentiment-analysis-in-python-with-vader-no-training-required-4bc6a21e87b8 (Accessed November 30, 2022).

Kuzminykh, Natalia (n.d.) Sentiment Analysis in Python With TextBlob. StackAbuse: https://stackabuse.com/sentiment-analysis-in-python-with-textblob/ (Accessed November 30, 2022).

MANS Media. (2022, November 30). NEWSPAPER DATA -- THE POWER OF NEWSPAPERS, PRINT AND DIGITAL. https://mansimedia.com/expertise/newspaper-data/ (Accessed November 30, 2022).

Office of the United States Trade Representative – About US – Benefits of Trade. https://ustr.gov/about-us/benefits-trade (Accessed November 30, 2022).

References

Selvaraj, Natassha. (2020, September 12). A Beginner’s Guide to Sentiment Analysis with Python. https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6 (Accessed November 30, 2022).

Wikipedia Foundation. (2022a, November 28). *S&P 500.* Wikipedia. https://en.wikipedia.org/wiki/S%26P_500 (Accessed November 30, 2022).

Wikipedia Foundation. (2022b, November 27). *The Wall Street Journal.* Wikipedia. https://en.wikipedia.org/wiki/The_Wall_Street_Journal (Accessed November 30, 2022).

Data Sources:

S&P 500 Stock Index Data. https://finance.yahoo.com/quote/%5EGSPC/history/ (Accessed November 30, 2022).

The Wall Street Journal. https://www.wsj.com/news/archive/years (Accessed November 30, 2022).