– Load the dataset, view basic structure, and check for null values.

Actually I tried to fetch the sec filing of the apple company from the api but not getting satisfied result many values are missing and also taking time . actually there is only 4 document per year so that I decided to extract data manually

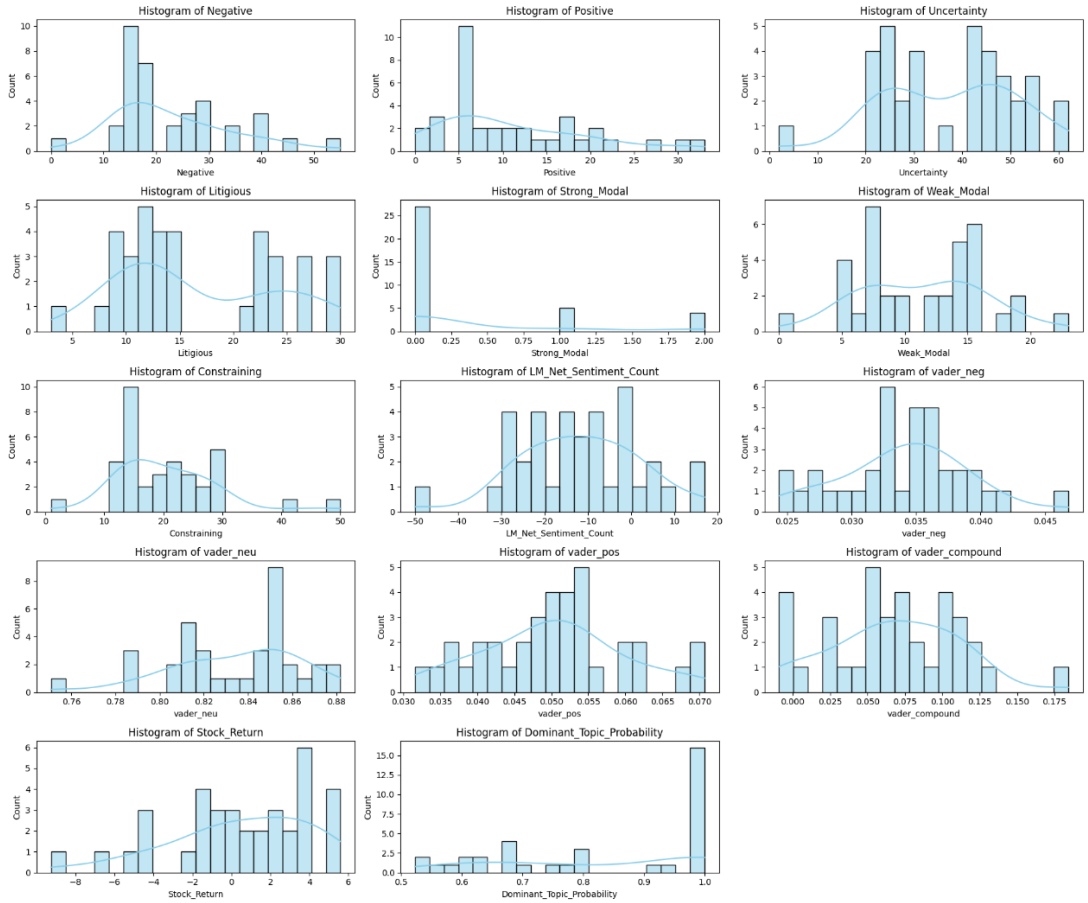
– Handle missing values, duplicates, and inconsistent formats.

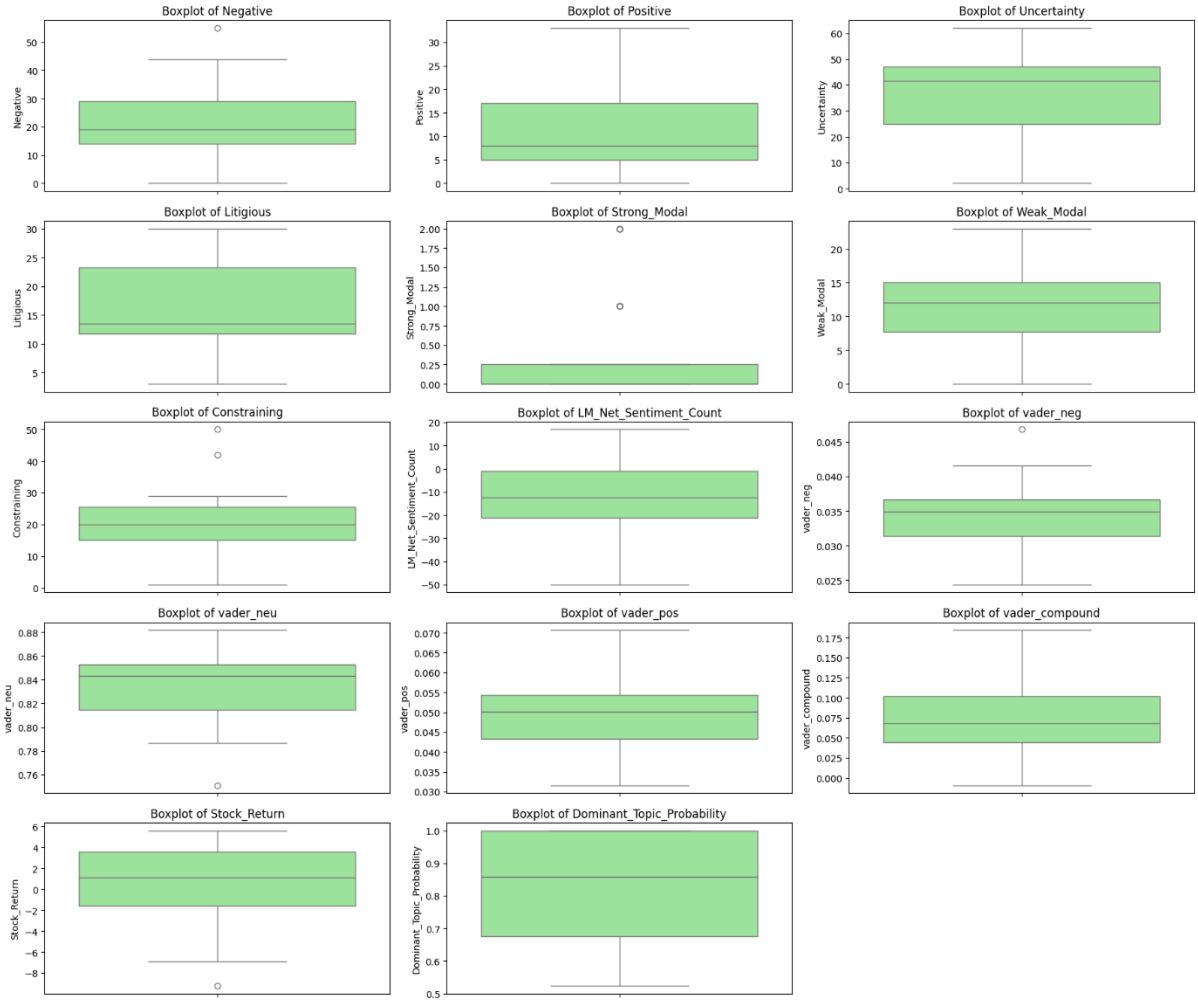
I check in my dataset I don't find this type of any issue

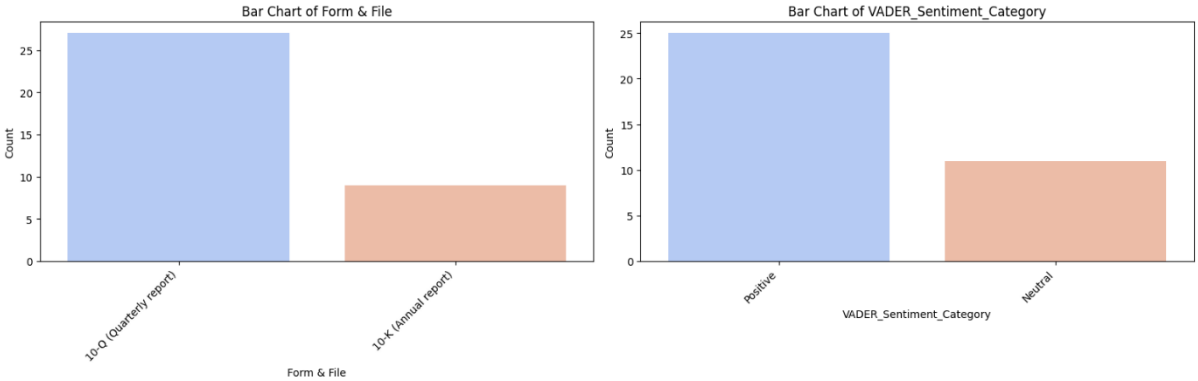
– Generate descriptive statistics for numerical and categorical features.

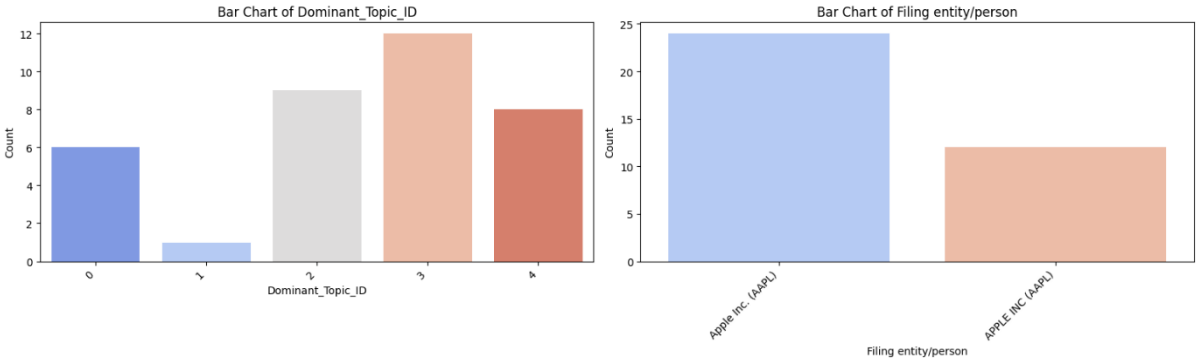
	word_count	sentence_count	char_length
count	36.000000	36.000000	36.000000
mean	3973.666667	109.944444	25041.083333
std	1289.543330	36.488702	8120.848676
min	416.000000	5.000000	2520.000000
25%	2820.000000	78.000000	17687.500000
50%	4337.500000	109.500000	27404.000000
75%	5147.500000	144.000000	32766.000000
max	5357.000000	162.000000	32767.000000

– Explore distributions of individual features using plots (histogram, boxplot, bar chart).

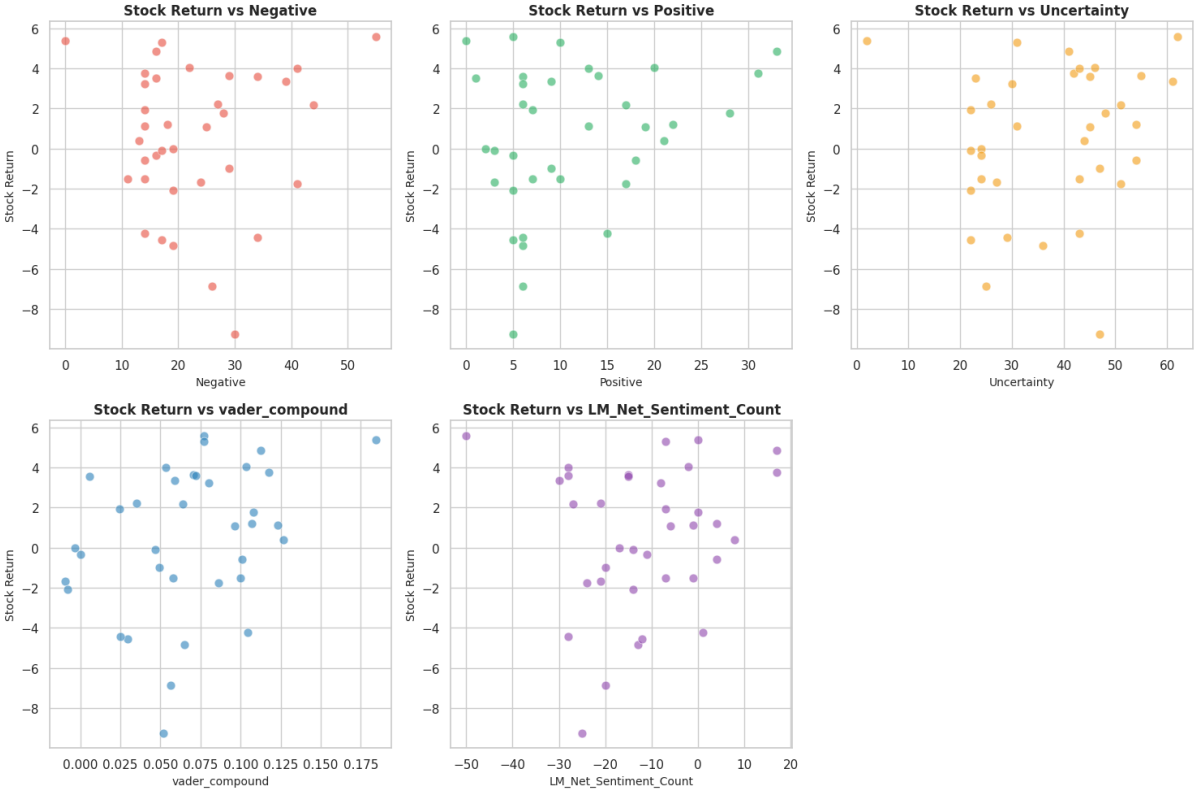


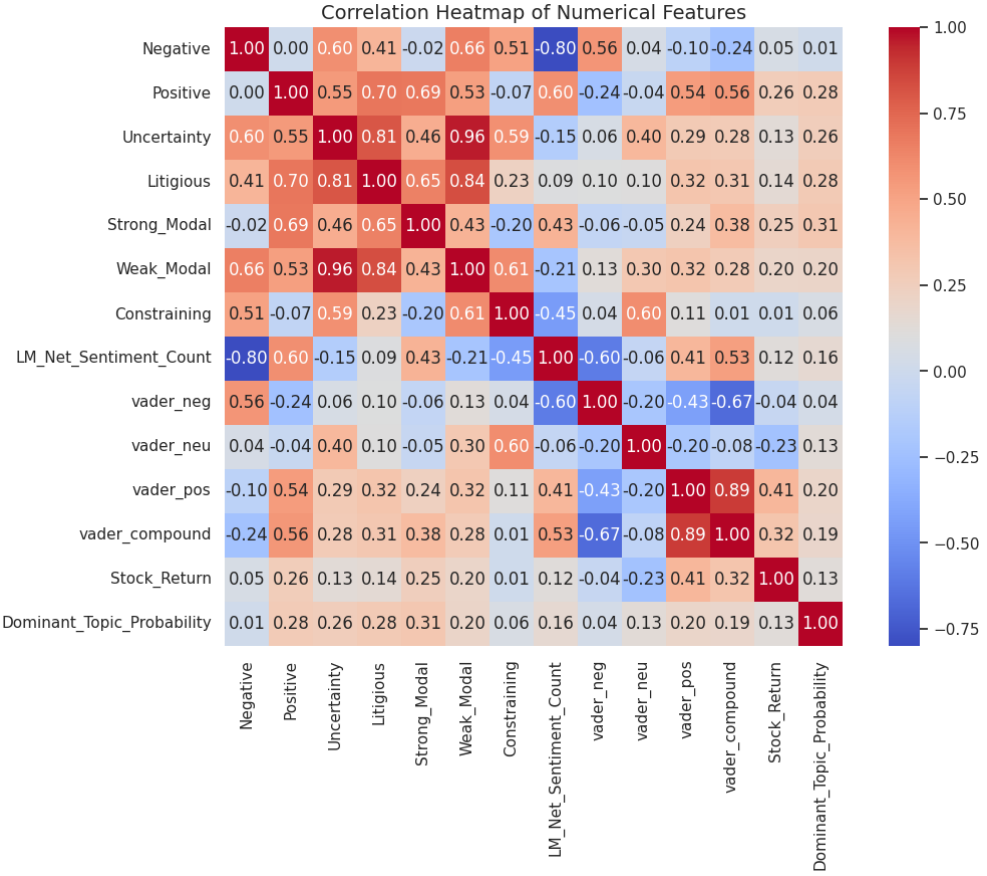

Boxplot of Negative
Negative
Boxplot of Positive
Positive
Boxplot of Uncertainty
Uncertainty
Boxplot of Litigious
Litigious
Boxplot of Strong_Modal
Strong_Modal
Boxplot of Weak_Modal
Weak_Modal
Boxplot of Constraining
Constraining
Boxplot of LM_Net_Sentiment_Count
LM_Net_Sentiment_Count
Boxplot of vader_neg
vader_neg
Boxplot of vader_neu
vader_neu
Boxplot of vader_pos
vader_pos
Boxplot of vader_compound
vader_compound
Boxplot of Stock_Return
Stock_Return
Boxplot of Dominant_Topic_Probability
Dominant_Topic_Probability


Bar Chart of Form & File
Count
10-Q (Quarterly report)
10-K (Annual report)
Form & File
Bar Chart of VADER_Sentiment_Category
Count
Positive
Neutral
VADER_Sentiment_Category

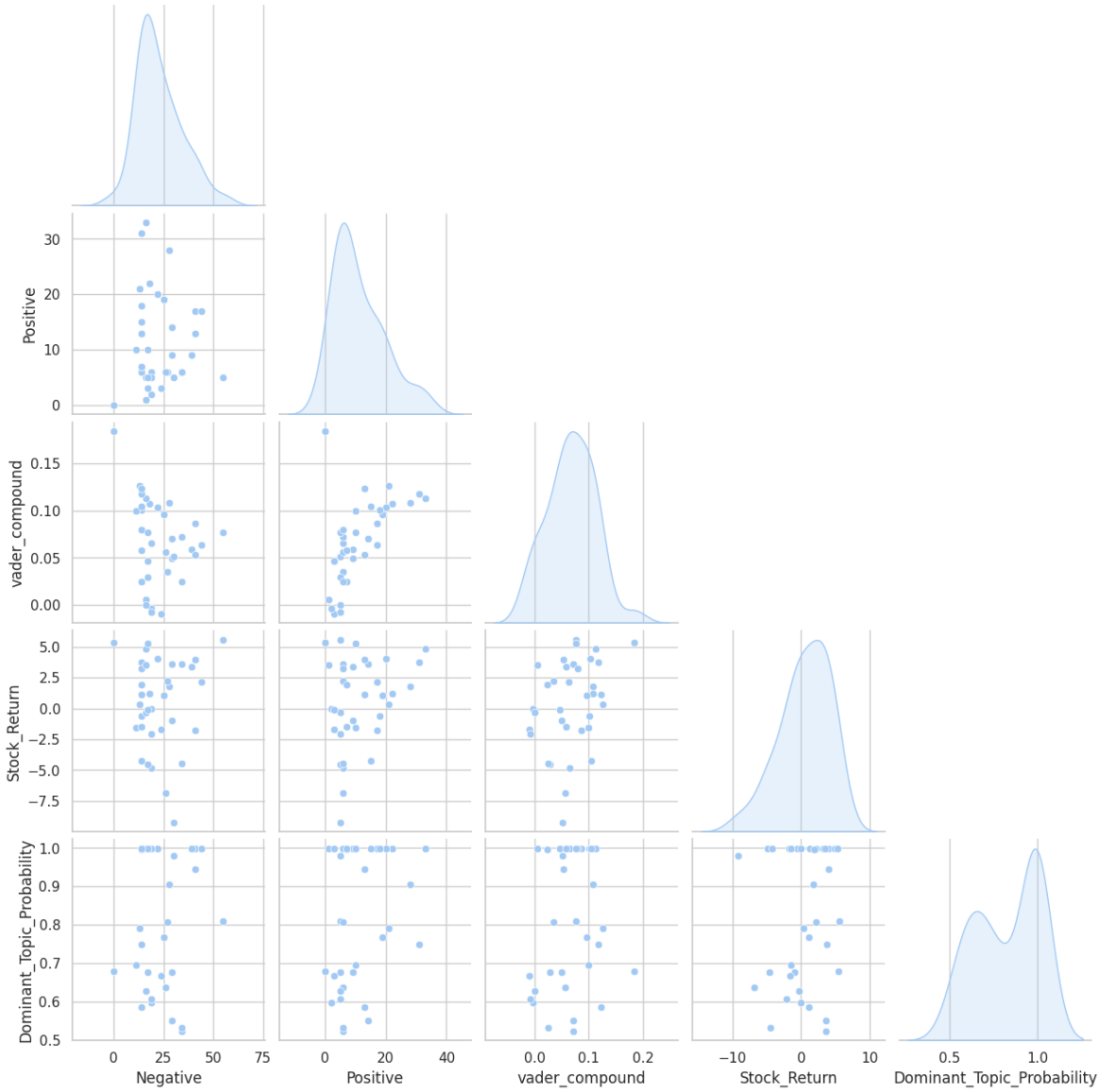

Bar Chart of Dominant_Topic_ID
Count
0
1
2
3
4
Dominant_Topic_ID
Bar Chart of Filing entity/person
Count
Apple Inc. (AAPL)
APPLE INC (AAPL)
Filing entity/person

– Study relationships between variables using scatter plots, heatmaps, and pairplots.

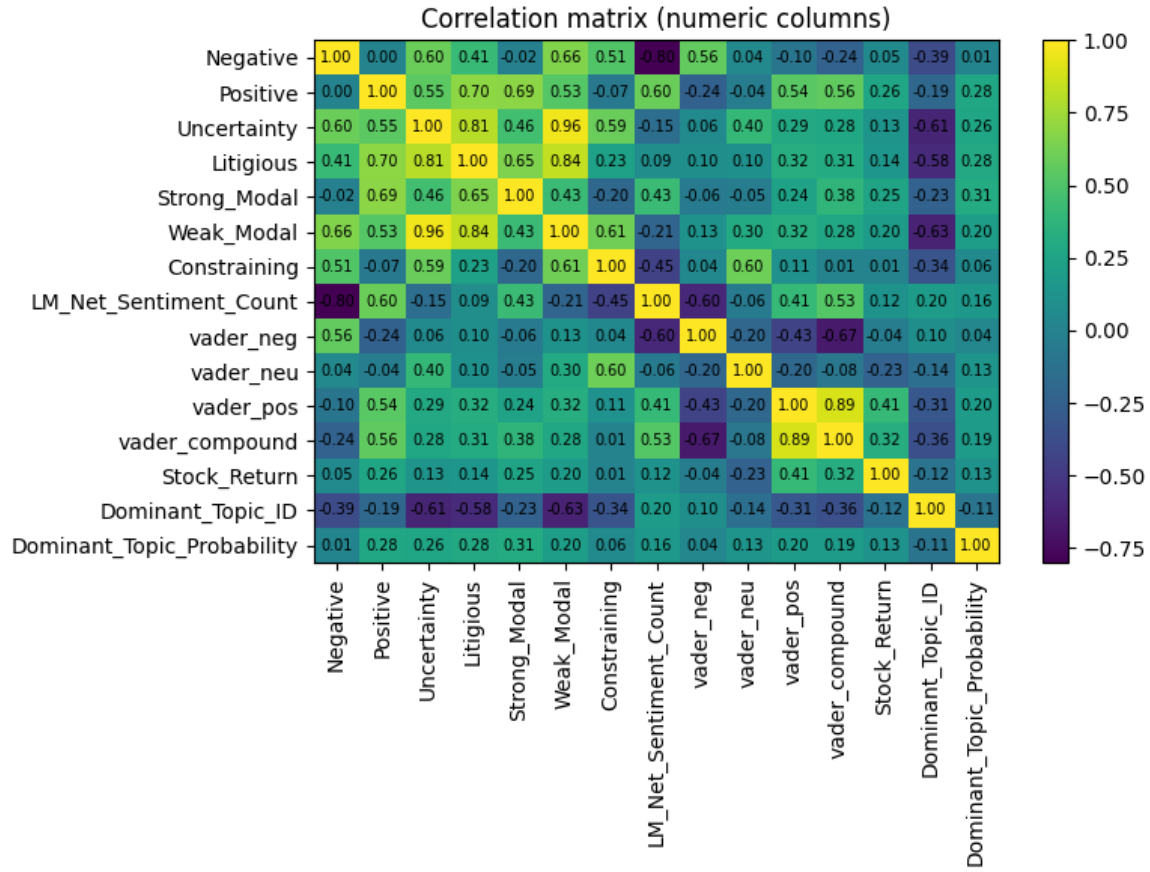




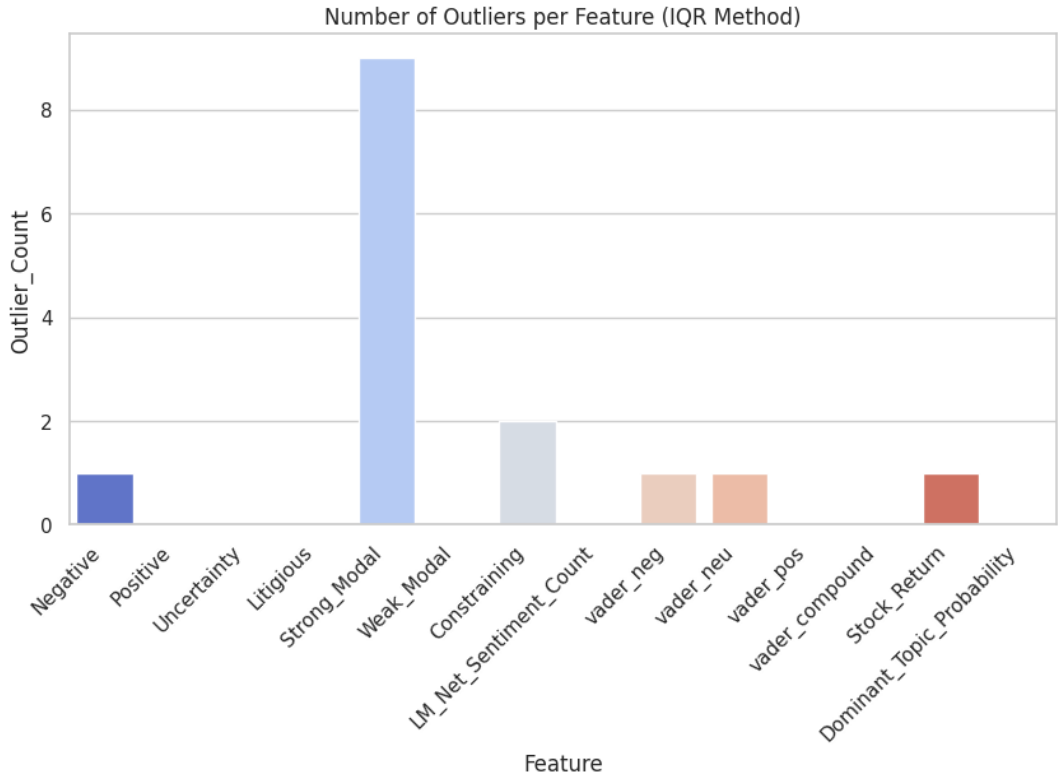
Pairplot of Sentiment and Stock Return Relationships

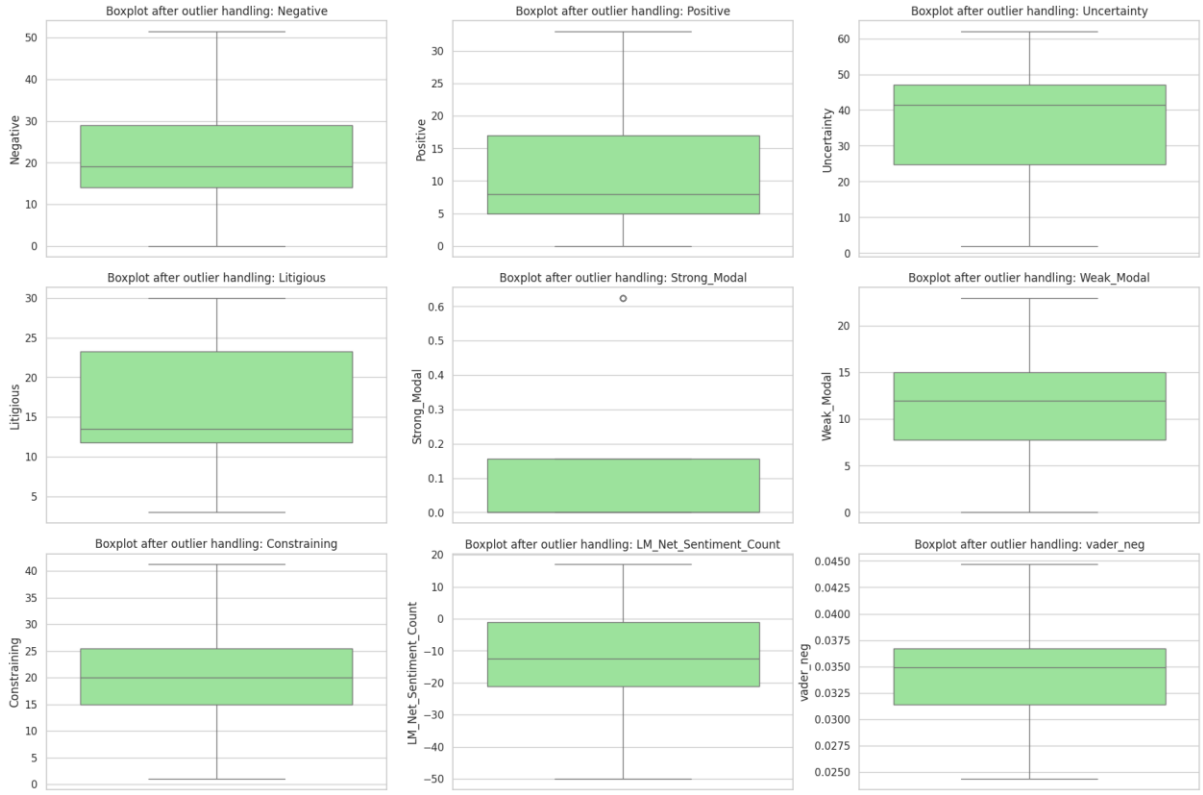


– Compute and visualize correlations to find dependencies.



– Identify and handle outliers appropriately.





– Create or transform features if relevant.

I created a some column :

“LM_Net_Sentiment_Count”: which tell the net count of the (positive - negative)

“Stock_Return”: I take a data from the yfinance ,after 30 day from the filing date

“Dominant_Topic_ID”:by using genism library using LDA model to analyze the word and put the similar in a one Topic ,I divided the whole Topic in to 5 category

“Dominant_Topic_Probability”: This column shows that LDA model is how much confident to put the paragraph in to any Topic

“Custom_Sentiment_Category”: based on the “vader_compound” and put threshold value like:(*Negative< 0.000 <Neutral<+0.05<Positive*)

– Summarize key findings using appropriate visualizations and a short written interpretation.

Correlation Between Numerical Features

- **High Positive Correlation (Sentiment/Tone):** There is a **very strong positive correlation** (close to 1.00) among the **LM (Loughran & McDonald) sentiment categories**: **Positive, Uncertainty, Litigious, Weak_Modal, and Constraining**. This suggests they often appear together or are measuring similar aspects of the text volume.

- **Negative Correlation:** The **LM_Net_Sentiment_Count** is **strongly negatively correlated** with both **Negative** and **Litigious** (-0.80 and -0.70). This makes sense, as a high negative or litigious count would decrease the net positive sentiment.
- **VADER Scores: vader_compound** shows a **moderate positive correlation** with **vader_pos** (0.89) and a **strong negative correlation** with **vader_neg** (-0.67), which is expected as the compound score is derived from the other VADER scores.

B. Correlation with Stock_Return

- **Low Correlation with Stock_Return:** It has a **very weak linear relationship** with almost all the sentiment and linguistic features. The strongest correlations are very low:
    - **vader_compound:** 0.32 (Positive)
    - **vader_pos:** 0.41 (Positive)
    - **LM_Net_Sentiment_Count:** 0.41 (Positive)
    - **vader_neu:** -0.23 (Negative)
- The scatter plots in image visually confirm this, showing no clear linear patterns between **Stock_Return** and key sentiment features like **Negative**, **Positive**, **Uncertainty**, **vader_compound**, or **LM_Net_Sentiment_Count**.

C. Correlation Between Textual Metrics

- **High Redundancy: word_count, sentence_count, and char_length** are **extremely highly correlated** (correlations of 0.98 to 1.00). This indicates they essentially measure the same thing—the length of the text—and one of them might be sufficient for most modeling purposes.

Key Takeaways for Modeling

1. **Strong_Modal Outliers:** The large number of outliers in **Strong_Modal**
2. **Multicollinearity:** The extremely high correlation among several LM features (Positive, Uncertainty, Litigious, etc.) and the textual metrics suggests **high multicollinearity**. Consider applying dimensionality reduction (like PCA) or removing some of these redundant features to stabilize model training.
3. **Weak Predictors:** The very weak linear relationship between most features and **Stock_Return** implies that a linear model might not be the best approach, or that a more complex feature engineering/model selection process is required to capture any non-linear relationship.

- A Jupyter Notebook :- https://colab.research.google.com/drive/16iVq1eGkkwKrILczZVyCScIPHBCC52hZ?usp=sharing