NLP and Factor Modeling of US Stocks

04.23.2024

Team 14

Professor: John Miller

Maria-Nefeli Karytinou

Hongbo Li

Peter Penev

Yifan Tao

Introduction

Our project aims to incorporate NLP and factor modeling to investigate three major sectors in the United states, the technology sector, the financial sector, and the healthcare sector. We focused on the ten largest stocks by market capitalization within each sector, and then we used different methods to try to predict the two day returns of the stock based on the sentiments of the articles posted on the specific day. Our goal is to examine whether the sentiments expressed in articles can influence stock returns and if this could be a crucial factor for trading, with the key metric we used being the correlation coefficient to determine the relationship between the sentiment and the returns. We also sorted the data by magnitude of sentiment, and by different ways to process an article with multiple companies being mentioned to attempt to find the best method that has the highest tangible results.

Data Collection

The project collects data from two principal sources: Yahoo Finance for numerical data and GNEWS API for text data.

Numerical data encompasses daily metrics such as opening and closing prices, and trading volumes for all companies. For the factor modeling, we collected data from 12/12/2013 to 03/05/2024 while for the NLP analysis, the data is limited to the period from 08/10/2020 to 02/04/2024, sourced from Yahoo Finance via its official APIs.

For text data, our datasets included 17,000 financial news covering all 30 companies from 2020 to February 2024 by utilizing GNews API. The dataset contains the publication date, title, description, full content, article's URL and the source of each article.

Data Assimilation and Preprocessing

The project is divided into two main phases: the NLP and the factor modeling. Each phase has its datasets, before processing to the final stage when both parts will be combined.

In the first stage (NLP) our dataset integration and preprocessing workflow is designed to prepare for sentiment analysis. Initially, we employed regular expressions to extract mentions of companies within the article's full content, linking these mentions with their respective stock tickers for subsequent matching with our numeric data.
To refine the dataset further, we undertook a minimal cleaning process to remove noise, such as irrelevant information from the beginning of articles (e.g. photo descriptions, or publication sources). This step aimed to enhance the accuracy of the sentiment analysis by facilitating the internal pre-processing of our sentiment analyzer.

Furthermore, we expanded this dataset to associate each article with a singular ticker symbol, resulting in duplicate articles when multiple companies were mentioned. Large proportion of our dataset consists of articles with more than 1 company.

Following this, we integrated the text dataset with the Yahoo finance data into one big dataset based on tickers and publication dates. We kept the numerical data (open price, closing price) only for days when articles were published, considering the publication day as day t. For each ticker, we included price data for day t, the preceding day t-1, and the following day t+1, and we further calculated the returns over one day (t to t+1) and two day (t-1 to t+1) windows. For the publication days that are holidays or weekends, we assigned as price on day t the next trading day.

This consolidated dataset will be the base of our analysis, including sentiment classification and topic modeling.

Section 1: Sentiment Analysis and Topic Modeling

In this stage, we aim to implement sentiment analysis and provide an analysis of the sentiment and the stocks' returns. Our goal is to find if there is a correlation between the news sentiments with the stock's returns in different sectors.

1.1. Sentiment Analysis

Sentiment analysis is part of NLP that focuses on identifying and quantifying the emotional tone within texts. It integrates NLP, data mining, ML, and computational linguistics to ascertain whether sentences, paragraphs, or documents express positive, negative, or neutral sentiments.

In the finance sector, sentiment analysis can be very useful for understanding market dynamics and evaluating investor and public sentiment, as well as economic indicators by analyzing financial news, social media posts, and financial reports. By leveraging such insights to forecast market movements and trends in specific stocks, sectors, or the whole economy, stakeholders and financial institutions are equipped to make well-informed investment decisions that rely on advanced analytical methods.

However, the complex and nuanced language of the financial domain poses some challenges to the extraction of meaningful sentiment information. Furthermore, due to the lack of large labeled financial datasets, it is difficult to utilize certain tools like neural networks to their full potential.

In our project, we aim to tackle this issue by leveraging FinBERT, a pre-trained large language model specifically designed for the financial sector. FinBERT is a specialized version of BERT model, fine-tuned for sentiment classification. It is built by further pre-training BERT using a subset of Reuters' TRC2, called TRC2-financial, comprising 1.8M news articles that were published by Reuters between 2008 and 2010. The resulting dataset includes 46,143 documents with over 29M words and nearly 400K sentences. In addition, Financial PhraseBank from Malo et al. 2014 was used for fine-tuning. Experimental results demonstrate that FinBERT outperforms other language models such as original BERT, VADER, and GPTs in sentiment analysis tasks within the financial sector.

1.2. Datasets

Using a pre-trained model without further fine-tuning, can introduce noise into our analysis due to lack of labeled data. To mitigate potential inaccuracies from the sentiment analysis we have segmented our dataset into three distinct datasets for experimental purposes.

Dataset 1: consists of articles that refer to a single company (7,405 articles)

Dataset 2: consists of articles that mention multiple companies but keep as a stock reference the most frequently mentioned company (17,018 articles)

Dataset 3: consists of articles that have been duplicated to refer to one article per company (17,018 articles expanded in 40,587 rows)

More specifically, Dataset 1 provides the cleanest version of our dataset as there is a one-to-one association between articles and companies. Dataset 2, exploits all the articles we found, contributing to a more complete dataset. By excluding multiple companies that are less frequently mentioned within each article, we ensure that the overall sentiment score reflects the sentiment of the one most frequently mentioned company. Dataset 3, attempts to capture the sentiment of multiple companies within each article. Although we do not employ aspect-based sentiment analysis in this project, which would capture the sentiment specific to each company, we aim to minimize potential errors that could arise from an overall sentiment analysis by using FinBERT.

For Sentiment Analysis we extract the sentiment of both the title and the full content of the articles. FinBERT provides probabilities indicating how likely it is for each article to be positive, negative, or neutral. For each of our datasets, we applied a weighting to these probabilities to calculate sentiment scores ranging from -1 to +1. In particular, the process to derive the sentiment scores involved the following steps:

1. For article i, the Sentiment Si is calculated as:

$$Si = P(positive) \cdot (1) + P(neutral) \cdot (0) + P(negative) \cdot (-1) \; for \; i = 1,\ldots, N$$

2. For each company c and each date d, the average sentiment $Savg$ is calculated

$$Savg(c, d) = \sum_{i=1}^{N} Si(c, d)/N$$

Where:

N: total number of articles for company c on date d
$Si(c, d)$: sentiment score of the ith article for company c on date d

Specifically for Dataset 3, we performed an additional step of multiplying the sentiment score by the importance of each article, calculated as the percentage of mentions of each company divided by the total number of mentions. In that way, we aim to minimize the impact of companies that are rarely mentioned or that might exhibit a sentiment opposite from the one generally captured.

A score closer to 1 indicates a stronger positive overall sentiment for a company c on date d whereas a score closer to -1 indicates a stronger negative sentiment for a company c on date d. A score around 0 does not indicate the text is neutral in the conventional sense (lacking strong sentiment), instead suggests a balance between positive and negative sentiment scores, which are translated to somewhat unclear sentiment.

1.2.1. Dataset 1: Data with a single ticker per article

This dataset comprises 7,405 articles, each one associated with a single company. The distribution of these articles across different tickers is visualized in the following barplot. The distribution of articles per company, ordered by their market capitalization from largest to smallest, is shown below.

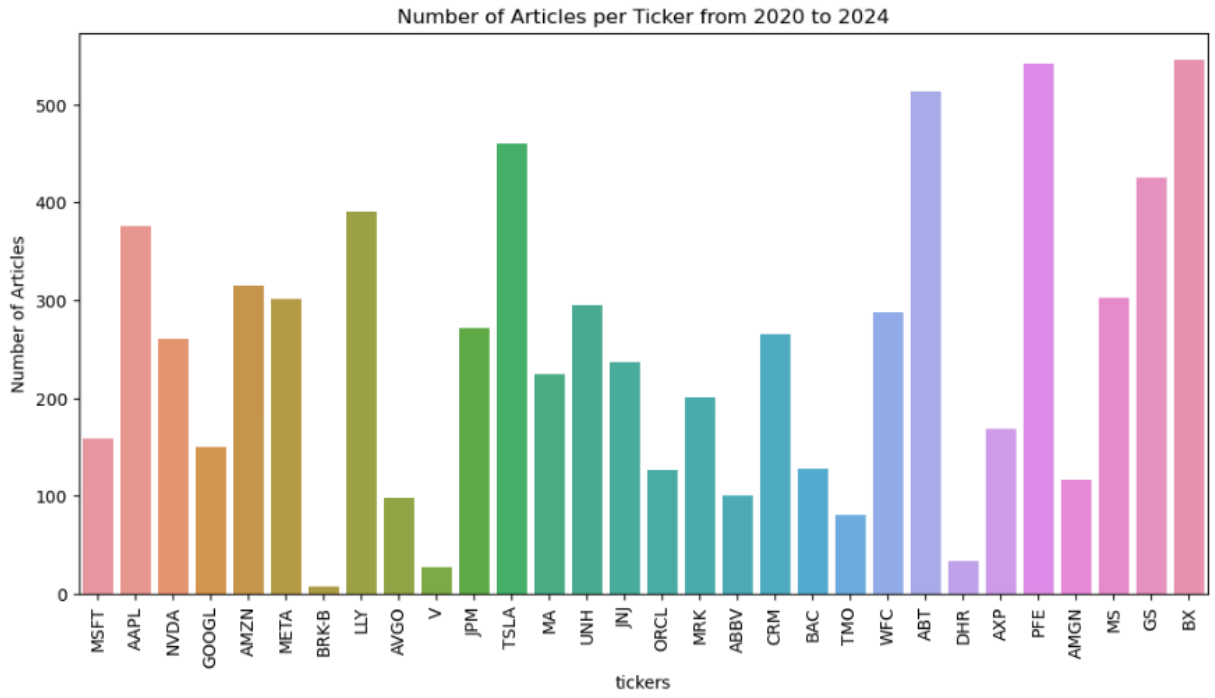


Fig 1.1.1: Number of articles per stock's ticker ordered by companies' market cap

By performing sentiment analysis on our dataset, using FinBERT we get the following:

Sentiment	*Title*	*Content*
Positive	1656	1464
Neutral	3750	3221
Negative	1999	2720

As mentioned in section 1.2, we standardize the sentiment scores on a scale from -1 to +1 for both the content and title. We also compute the average sentiment for the title and content and then we remove from the dataset duplicate entries to get average sentiments per ticker for each date on which at least one article was published. Therefore, the dataset now consists of 4,076 rows. Following this, we categorize the scaled sentiment scores into four intervals that correspond to distinct categories:

Negative for scores in [-1, -0.5]
Somewhat Negative for scores in [-1, -0.5]
Somewhat Positive for scores in [0, 0.5] and
Positive for scores in [0.5,1]

The distribution of articles fitting in each category for content and title sentiments is presented below:

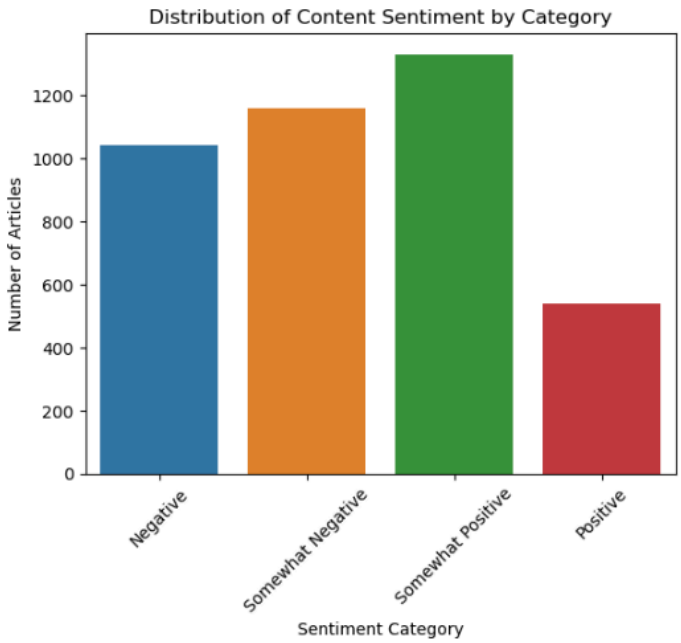


Fig 1.1.2: Number of articles per content category

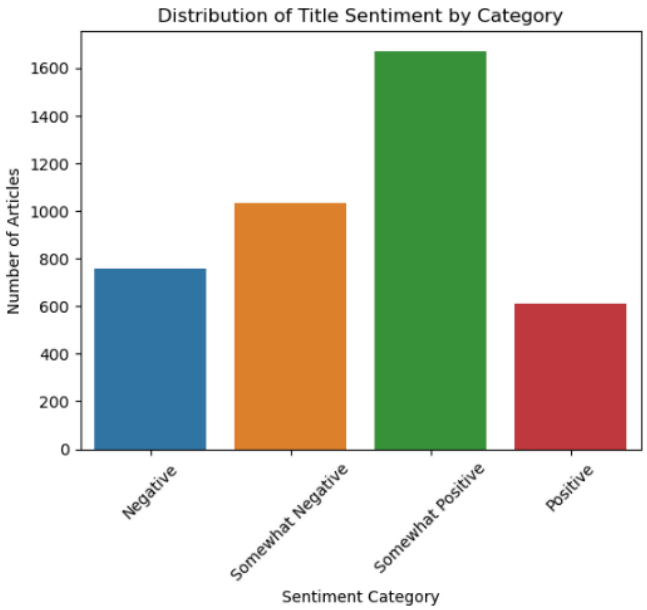


Fig 1.1.3: Number of articles per title category

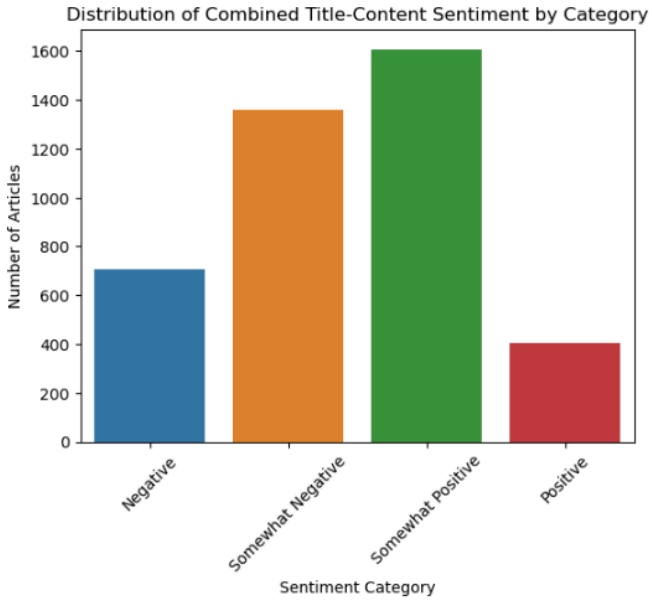


Fig 1.1.4: Number of articles per combined title-content sentiment category

Our goal is to explore the correlation between news and returns. Therefore, we examine the two-day period returns, from day t-1 - one day before the article's publication- to day t+1 - the day after the article's publication. To this end we compute the correlation between our dataset's sentiment scores and the two-day returns for the full content, the titles and the combination of title-content. The correlation matrix shows higher correlation values of 0.085 for the combined title and content sentiment, 0.77 for the content, and a slightly lower correlation of 0.71 for the title. While these correlations might initially appear low, they suggest a potential linear relationship between sentiment scores and immediate market performance.

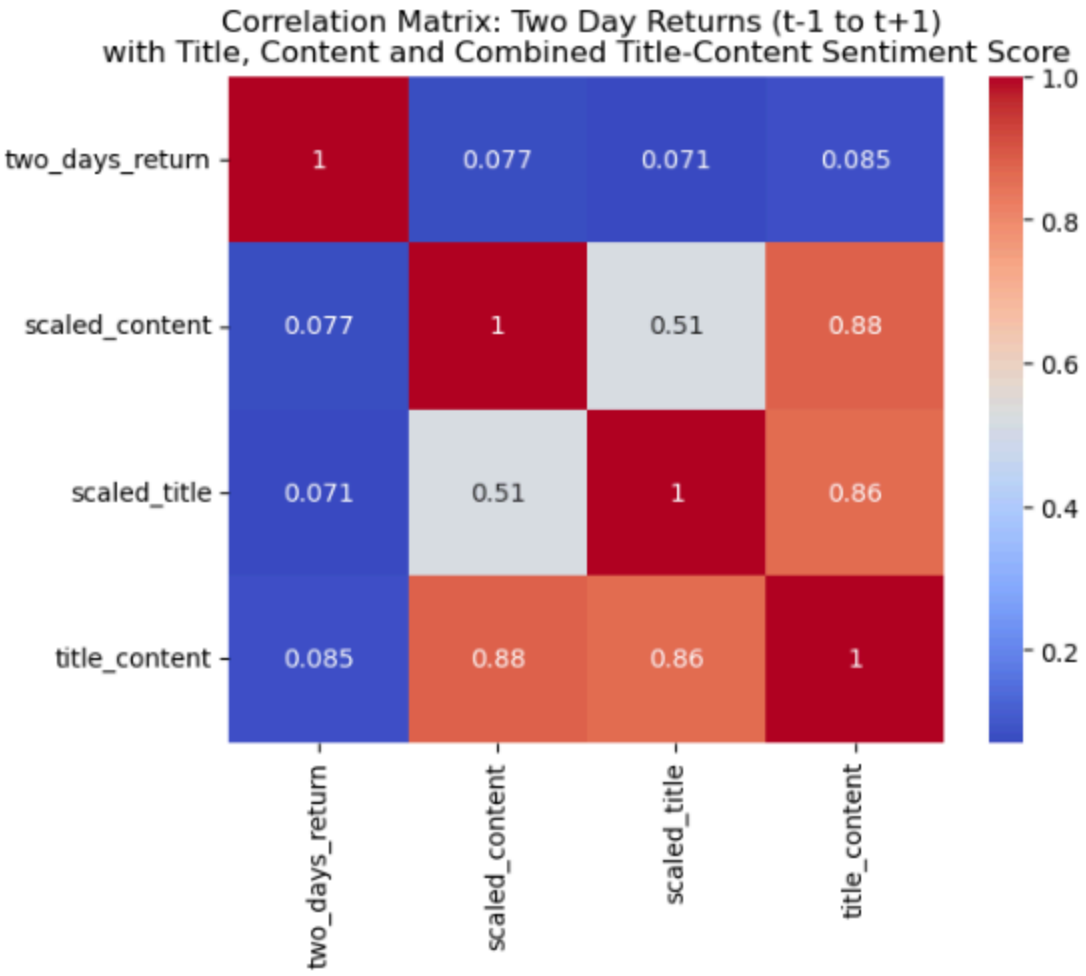


Fig 1.1.5. Correlation Matrix of Two-Day Returns with Title, Content, and the Average of Title and Content Sentiment Scores

Following the overall sentiment correlation analysis, we delve deeper into the sector-specific dynamics. We have categorized each article according to the sector of the company discussed in it. From Fig 1.1.6 it appears that content and the combined title-content sentiment have slightly higher correlations with two-day returns comparatively to news title sentiments in the technology and financials sectors. In contrast, for the healthcare sector, the trend is reversed. The sentiment of news titles presents a stronger correlation with returns while also achieving the highest correlation among all sectors and sentiment types examined. This suggests that technology-related articles and health care headlines may have a stronger influence on their respective sector's returns within the two-day window.
Overall, the data indicates that there is a small sector-based variation between sentiment scores and short-term stock returns.

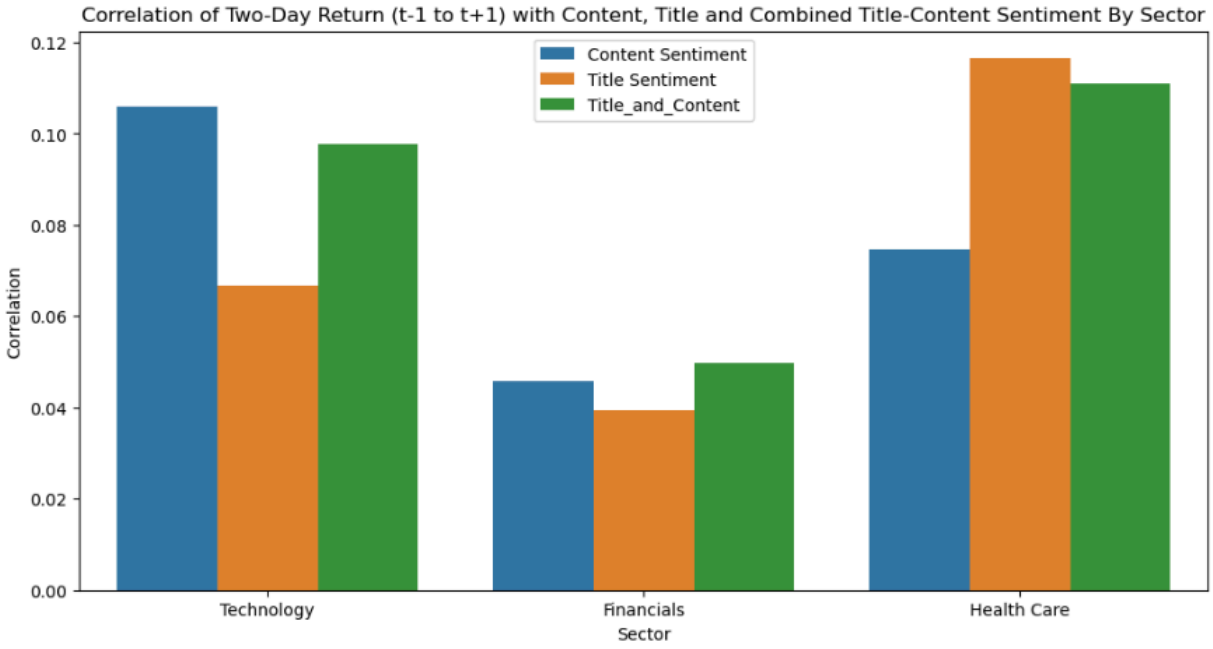


Fig 1.1.6: Correlation of Two-Day Returns (t-1 to t+1) with Content, Title and the Combined Title-Content Sentiment by Sector

These variations can be shown in the following scatterplots:

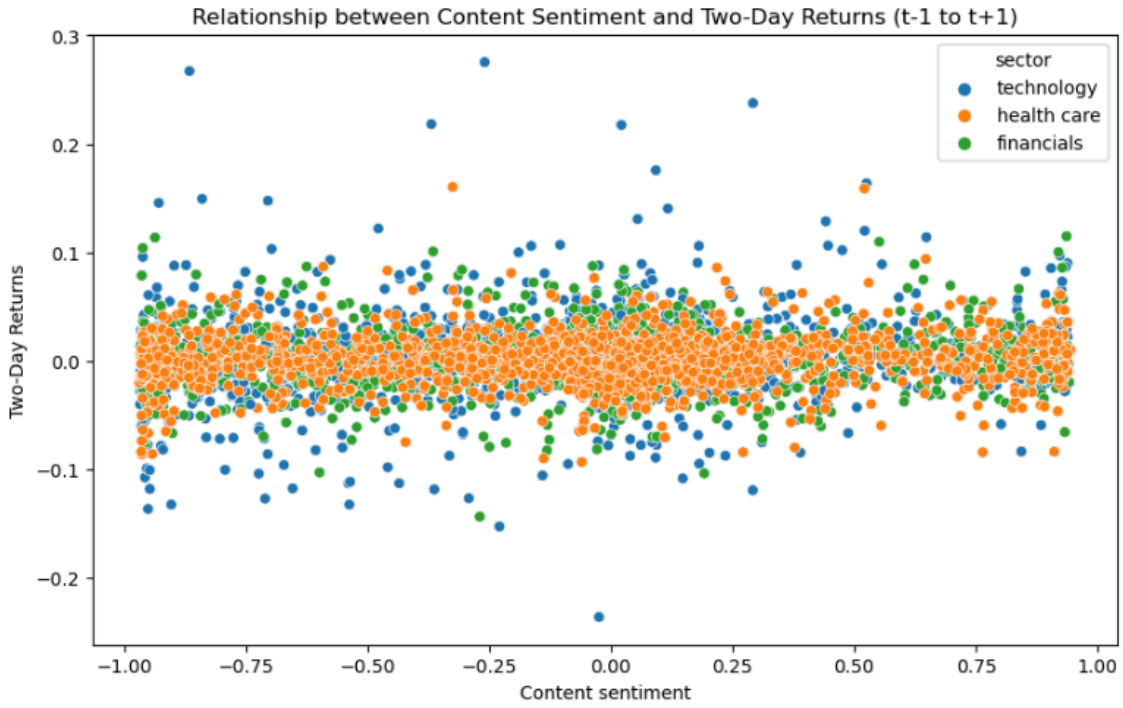


Fig 1.1.7. : Scatterplot of Content Sentiment score vs Two-Day Returns

At first glance, there does not appear to be a strong correlation between content sentiment and two-day returns for any sector. There is substantial overlap in the data points between different sectors suggesting that the range of returns and sentiment is quite similar. However, the technology sector exhibits a greater degree of variability, including more extreme values than other sectors on both positive and negative ends. This spread in the data points for the technology sector might indicate a greater sensitivity to the sentiment conveyed in news content.

Moreover, we observe the same behavior for the relationship between title sentiment and two-day returns. In addition, there is a dense aggregation of data points around the center of the sentiment axis approximately for values ranging from -0.25 to 0.25 suggesting a more neutral overall sentiment related to the titles. Similarly, for the title-content average, as depicted in Fig:1.1.8 there is no difference from the previous patterns observed.

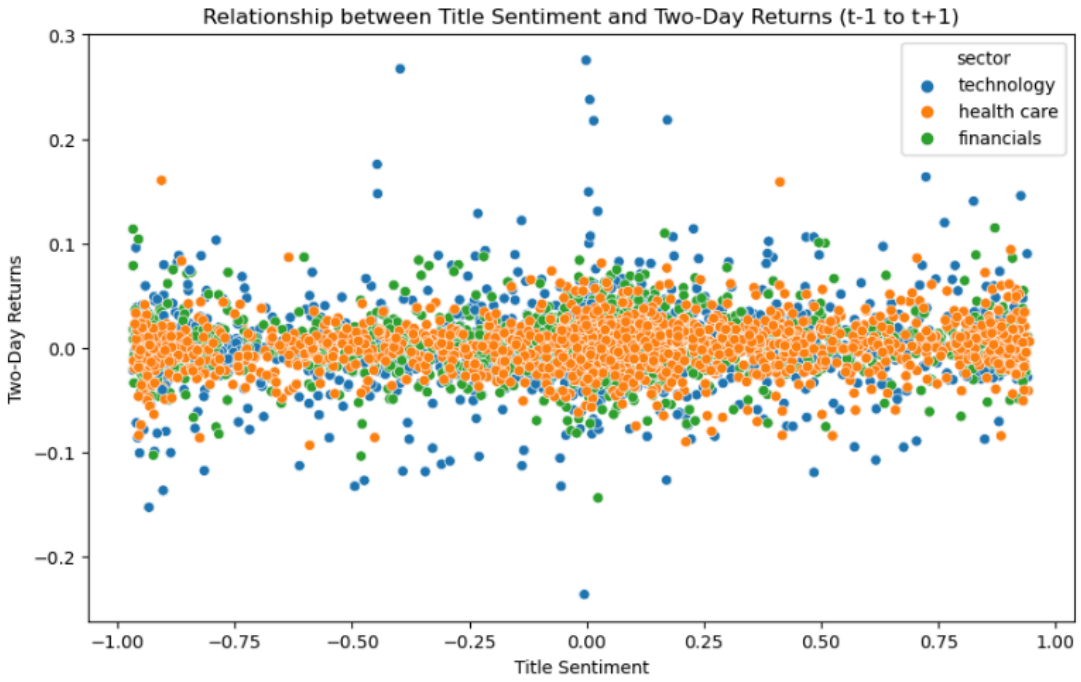


Fig 1.1.8. : Scatterplot of Title Sentiment Score vs Two-Day Returns

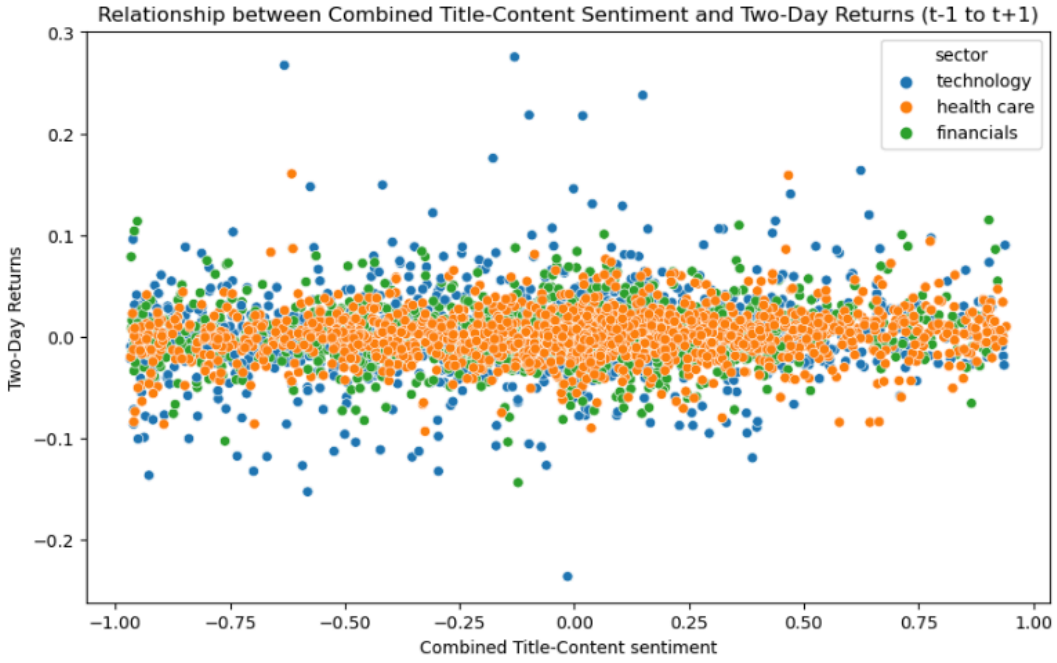


Fig 1.1.9. : Scatterplot of Combined Title-Content Sentiment Score vs Two-Day Returns

By applying a Pearson correlation test we find that the correlation between sentiment and two-day returns, for the content, the title, and the combined title-content sentiment, is statistically significant for the technology and healthcare sectors.

The results are presented in the following table:

		Sector		
		Technology	Financials	Health Care
Content	Correlation	0.106	0.046	0.075
	P-value	0.000057	0.116	0.0043
Title	Correlation	0.067	0.039	0.117
	P-value	0.011	0.177	0.000008
Title-Content	Correlation	0.097	0.049	0.11
	P-value	0.0002	0.087	0.000022

Table 1.1.1: Pearson Correlation Test: Correlation and p-values for content, title and combined title-content sentiment with two-day returns for each sector

While the Pearson correlation test underlines the significance of sentiment's impact on stock returns in each sector, we enrich our exploratory data analysis by visualizing the distribution of two-day returns within each sentiment category.

From the violin plot (Fig 1.1.11), we observe that for each sentiment category corresponding to content sentiments, the majority of the two-day returns are clustered around zero, which aligns with the expectation that daily returns often show small fluctuations. The median return for each category, indicated by the white dot, also appears to be near zero, suggesting no substantial median gain or loss over the two-day period regardless of sentiment.
For the "Positive" sentiment category, the range of returns is skewed towards more positive values. This could imply a tendency for clear positive sentiment to correlate with improved stock performance.

On the other hand, the "Negative" sentiment category presents an unexpected behavior with positive outliers. These outliers suggest cases where stocks had positive returns despite negative sentiment. This indicates that other factors may influence stock returns beyond content sentiment.

In addition, for "Somewhat Negative" sentiments, there is a wide range of stock returns, indicating a more unpredictable effect for this group. This broadness suggests that the

market's reactions to subtle sentiment vary significantly, leading to various stock performance outcomes.

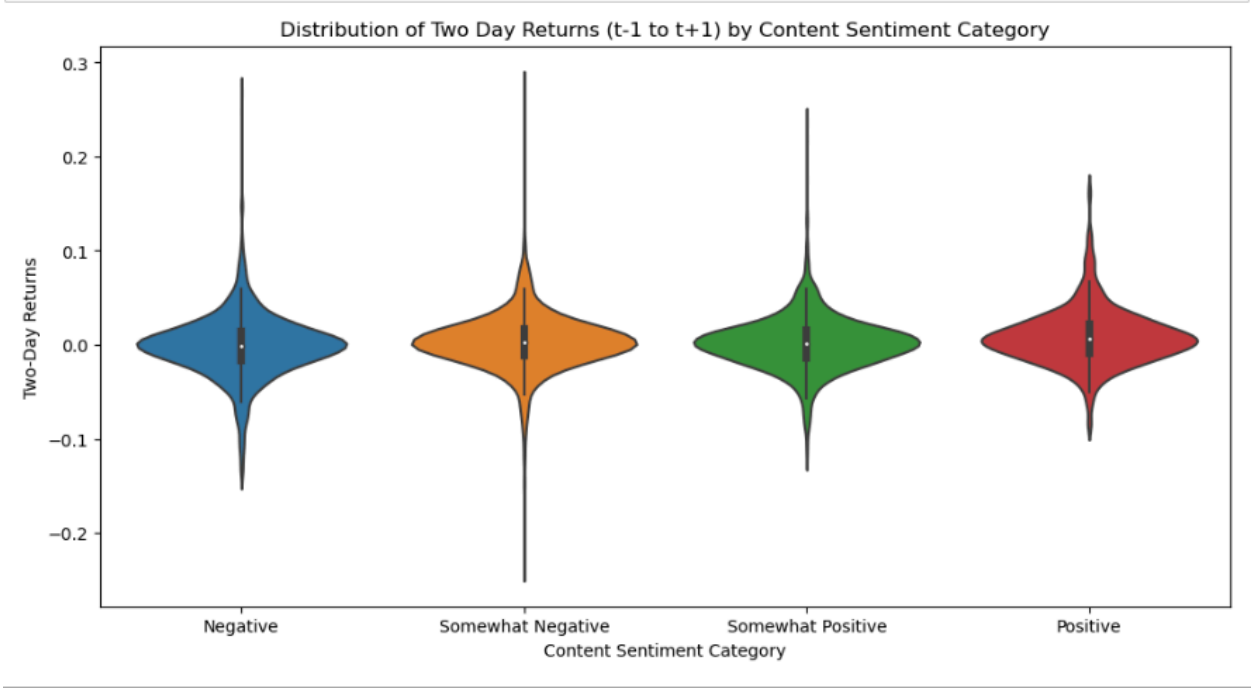


Fig 1.1.11. : Violinplot showing the Distribution of Two-Day Returns by Content Sentiment Category

Similarly, the following two violin plots for title and combined title-content sentiment, respectively, demonstrate almost identical patterns

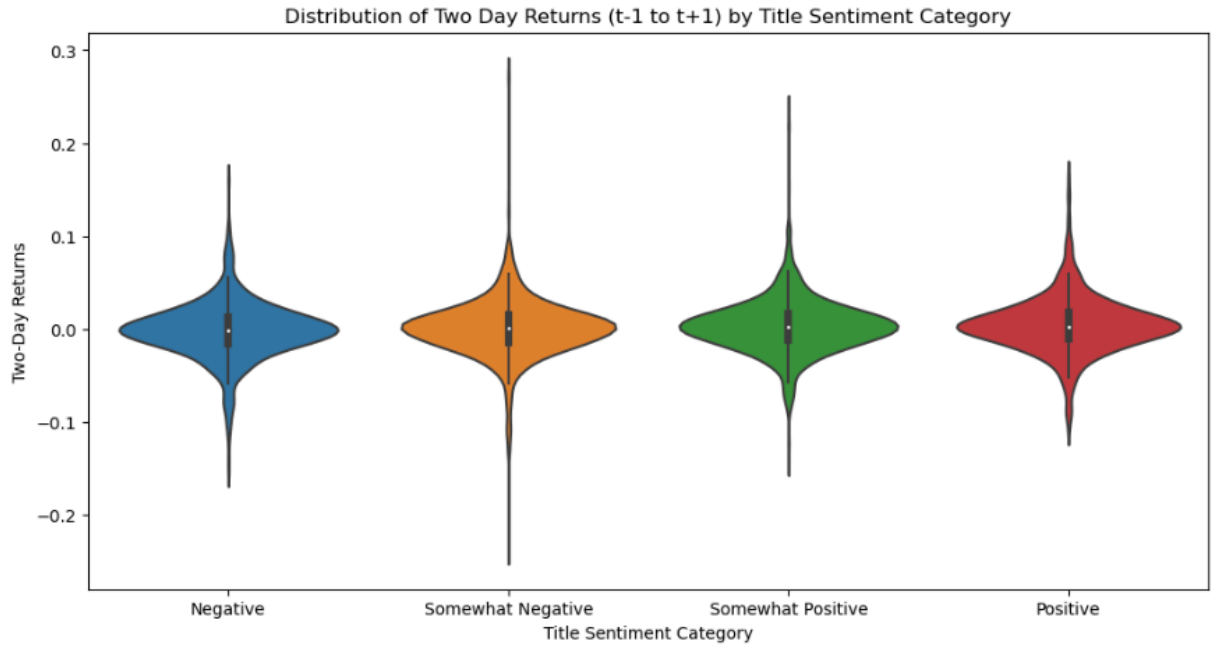


Fig 1.1.12. : Violinplot showing the Distribution of Two-Day Returns by Title Sentiment Category

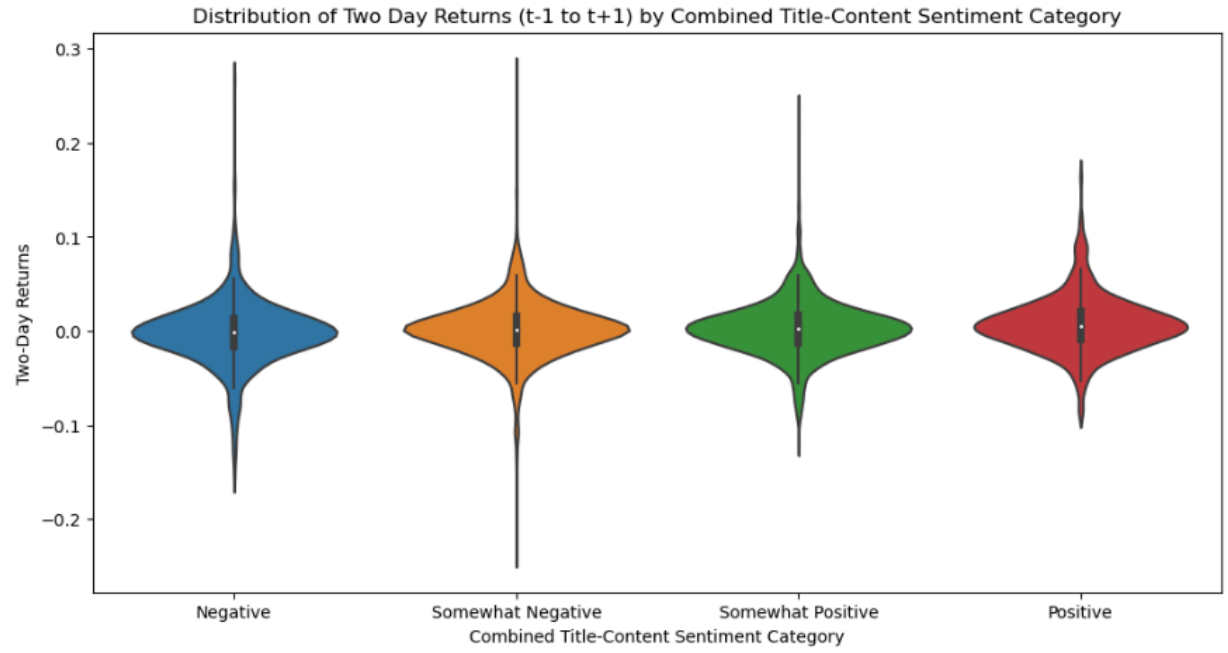


Fig 1.1.13. : Violinplot showing the Distribution of Two-Day Returns by Combined Title-Content Sentiment Category

To better understand the relationship of the sentiment in each category with the two-day returns, we further examine each sector individually.

The following violin plots (Fig 1.1.14) demonstrate similar patterns within each sector across all sentiment categories but with slight differences across sectors within each category. The longer whiskers on the technology sector's violins compared to the other sectors suggest that it experiences more extreme outliers in two-day returns. Additionally, the narrower appearance of these violins along the sentiment category axis indicates a more concentrated range of sentiment scores within each category, possibly due to less polarizing topics.

A particular point of interest is in the "Somewhat Negative" sentiment category.
Here, the return outliers vary across sectors, implying that when the overall sentiment is not clearly negative, it might be interpreted in different ways by the market. For example, the healthcare sector shows a slight skew towards positive returns whereas the financials show a negative skew. The technology sector maintains a balance in its distribution of positive and negative extremes, just like the overall behavior observed in the previous violin plot.

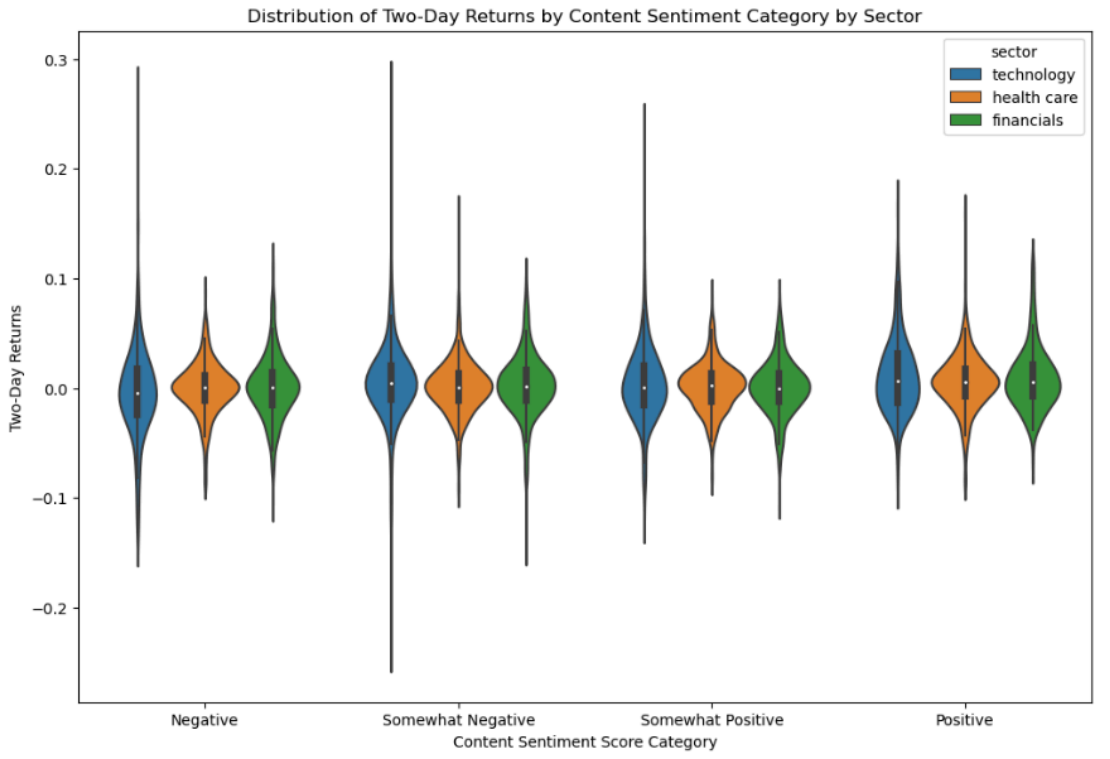


Fig 1.1.14: Distribution of Two-Day Returns by Content Sentiment for each Sector

When comparing the impact of title sentiment to content sentiment on two-day returns across sectors, a few distinct patterns emerge.

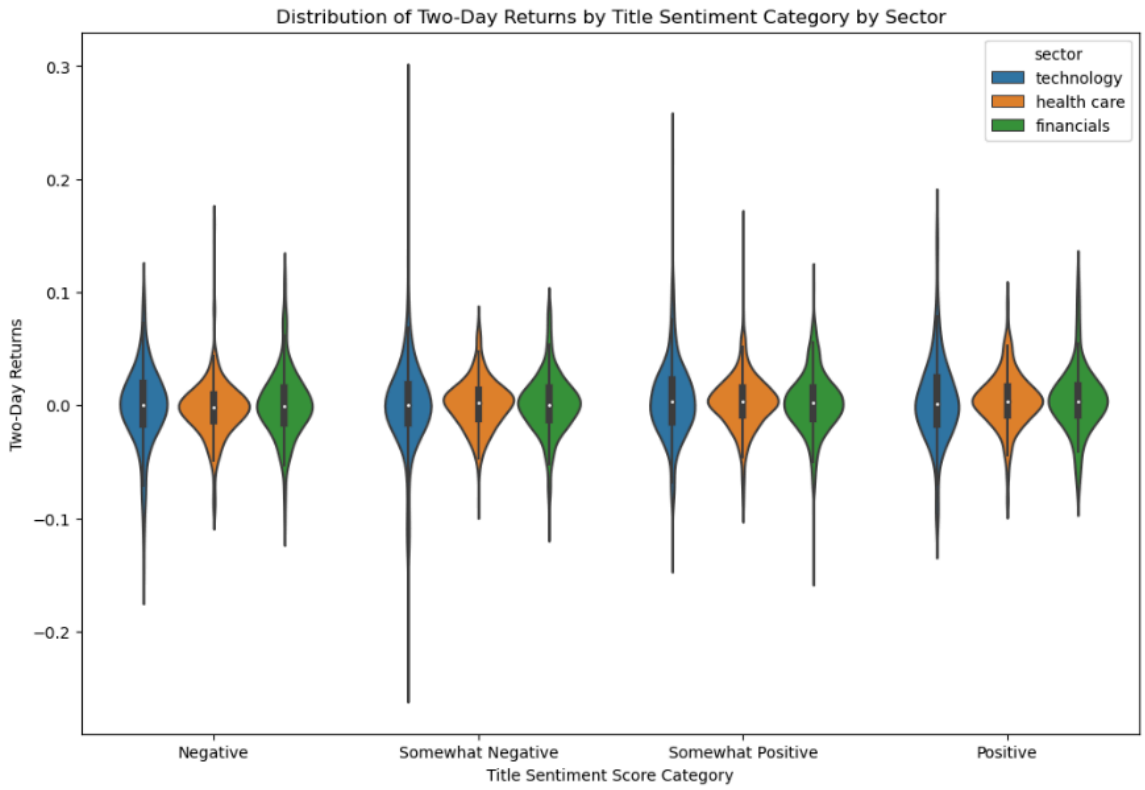


Fig 1.1.15: Distribution of Two-Day Returns by Title Sentiment for each Sector

For the technology sector, the “Negative” title sentiment category shows a distribution skewed towards negative returns, a contrast to its previously observed content sentiment behavior where the distribution appeared more positively skewed. This suggests that explicitly negative sentiments as reflected just from news titles is more likely to lead to negative returns in the technology sector than the ones related to the full content sentiment. Conversely, for the this category, in the healthcare sector, the previously observed symmetric pattern, where returns didn’t move far from zero, now has been shifted reflecting the presence of more positive outliers.

Moreover, the healthcare sector’s “Somewhat Negative” category appears narrower suggesting a concentration of more moderate returns in response to news titles. The distributions for the positive categories - Positive, Somewhat Positive - reveal the same patterns with the content-related sentiments. This shows that sentiment related to titles does not dramatically change the relationship between sentiment and two-day returns when the overall sentiment is positive or somewhat positive.

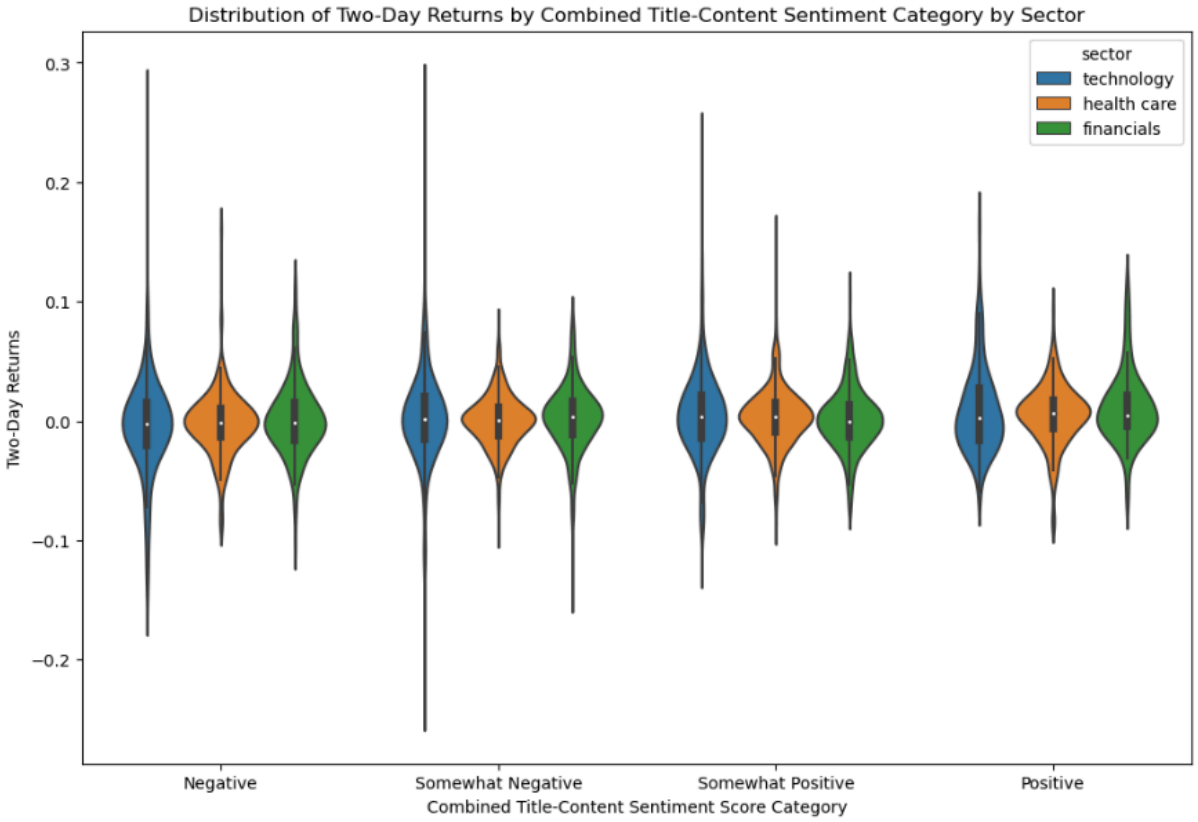


Fig 1.1.16: Distribution of Two-Day Returns by Combined Title-Content Sentiment for each Sector

Lastly, for the combined title-sentiment, we observe similar trends to both the content and title sentiment distributions across all sectors and all categories. For example, in the "Negative" category the health sector shows identical patterns with the title's sentiment distribution. Conversely, in the "Somewhat Negative" and "Somewhat Positive" categories the financial sector demonstrates similar patterns to the content sentiment distribution.

To facilitate a direct comparison of the average two-day returns across different sentiment categories and sectors for content-related articles the following barplot has been created.

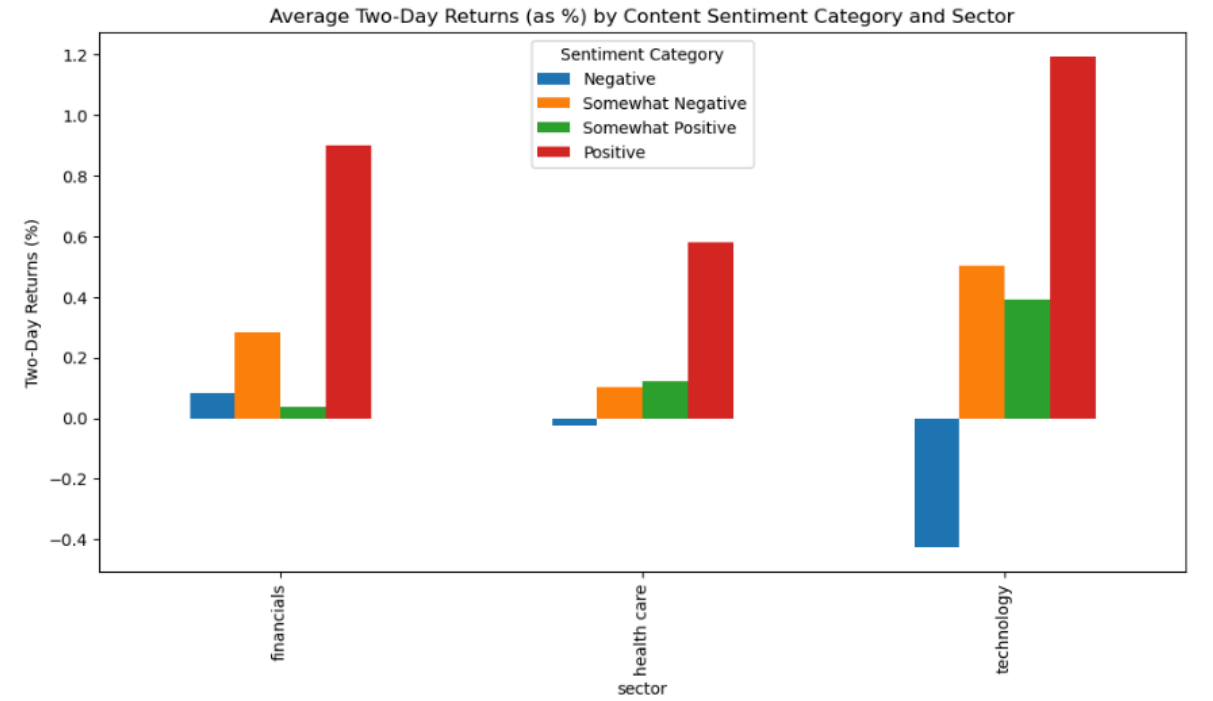


Fig 1.1.17: Average Two-Day Returns (as %) by Content Sentiment for each Sector

The financials sector exhibits positive average returns across all sentiment categories, suggesting that negative or somewhat negative sentiment does not necessarily lead to negative returns. However, positive sentiment is associated with the highest positive return, averaging 0.9%, indicating that the sector is highly responsive to clearly positive sentiment related to the article's content. Conversely, in the healthcare sector, we observe a relatively narrower range of average returns across all sentiment categories, with somewhat positive and somewhat negative sentiments yielding almost identical outcomes. An explanation could be that unclear sentiment can be perceived differently or overlooked by the market, leaving space for diversity in returns. Additionally, the technology sector stands out with highest positive and negative average returns for their respective categories. These pronounced results suggest a sector with higher sensitivity to clear-cut sentiments, either positive or negative. Furthermore, when comparing across categories, it appears that "Negative" content sentiment yields negative average returns for both the technology and healthcare sectors but not for financials. This pattern indicates that negative news sentiment might have a more direct impact on stock performance in the two of three categories. Interestingly, somewhat negative sentiment is associated with positive two-day returns across all sectors, with financials and technology experiencing relatively higher gains. These positive returns despite the slight negative sentiment could reflect a market tendency to undervalue or disregard ambiguous overall sentiments. Moreover, for "Somewhat Positive" sentiments there are slightly positive returns suggesting that mildly positive news may be related to positive returns but may not be a strong driver of stock

performance over such a short-term window. However, for clear positive content sentiment, all sectors show a robust average return, indicating a strong positive reaction to favorable sentiment news.

For the sentiments derived from the articles' titles, the following plot reveals a similar behavior to that of article content across all sectors for all categories. However, there are some distinctions, particularly in the smaller ranges of average returns, and on more pronounced effects in certain cases. More specifically, in the healthcare sector, negative sentiment appears to have a more substantial impact on returns, as reflected by more negative two-day returns on average. Moreover, the technology sector has smaller average return differences between positive and somewhat positive sentiments, implying that even mildly positive sentiments could have a positive impact on stocks' returns. The data is presented below:

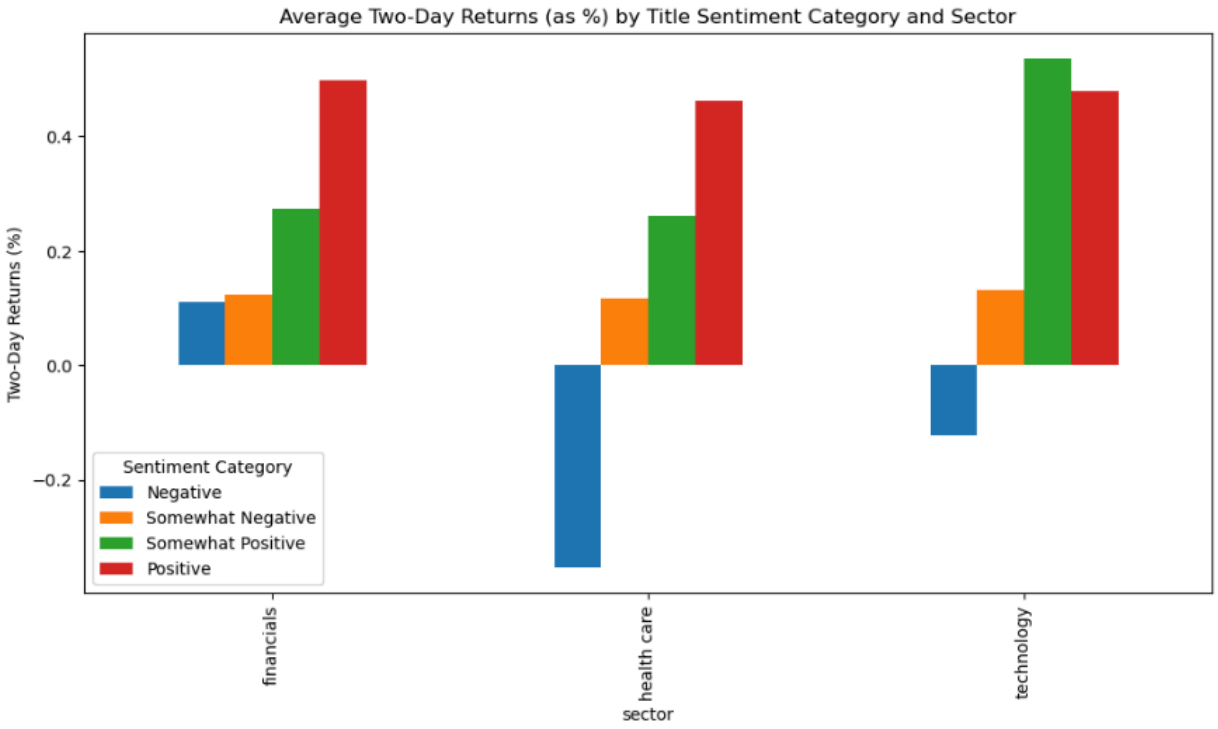


Fig 1.1.18: Average Two-Day Returns (as %) by Title Sentiment for each Sector

For the sentiments derived from the average of the articles' titles and content, the plot reveals a behavior similar to the title sentiment category. However, we observe different effects on certain categories across the sectors. In healthcare although negative sentiment leads to negative returns, the impact is less significant comparatively to the impact of title-sentiments in the same category. In contrast, the technology sector for negative sentiments exhibits more negative returns just like the pattern in content sentiment. As

expected, clearly positive sentiments are related to more positive two-day returns, while also the in-between categories (somewhat positive, somewhat negative) seem to have a lower but also positive association with the returns.

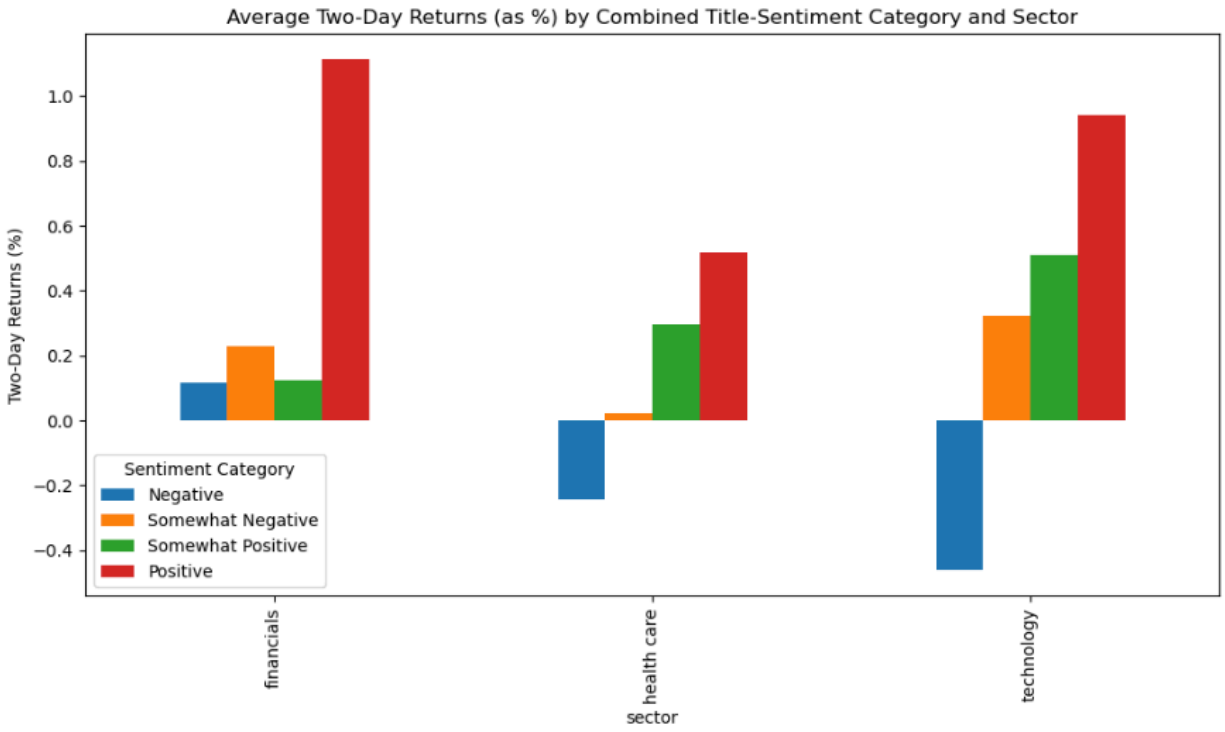


Fig 1.1.19: Average Two-Day Returns (as %) by Combined Title-Content Sentiment for each Sector

Conclusions

The analysis of the impact of sentiment derived from the titles, the content and their average of financial news and articles, on two-day stock returns reveals similar patterns across each type of sentiment but slight differences across sectors.

Overall, both title, content, and their combination show a statistically significant correlation with stocks' two-day returns in the technology and healthcare sectors, indicating there might be some sensitivity of these markets to news sentiment. Specifically, the technology sector shows a more significant response to both positive and negative sentiments. Similarly, the healthcare sector illustrates notable sensitivity particularly to title sentiment, indicating the importance of how news is presented in influencing stock performance. Conversely, the financials sector presents a unique pattern. Despite the sentiment, it often yields positive average returns, implying a different market perception of news, especially when related to negative sentiments.

Importantly, the analysis reveals that when sentiments are explicitly positive or negative generally lead to positive or negative average returns, respectively, in the technology and healthcare sectors. This suggests that the market reacts more rationally to news that leaves a clear positive or negative sentiment. However, in the financials sector, it appears that only distinctly positive sentiments significantly affect positive two-day returns.

1.2.2. Dataset 2: Data with the most frequently mentioned ticker for each article

Dataset 2 comprises all 17,018 articles each mentioning one or more companies. From sentiment analysis, we assign a sentiment score to the most frequently mentioned company in each article.

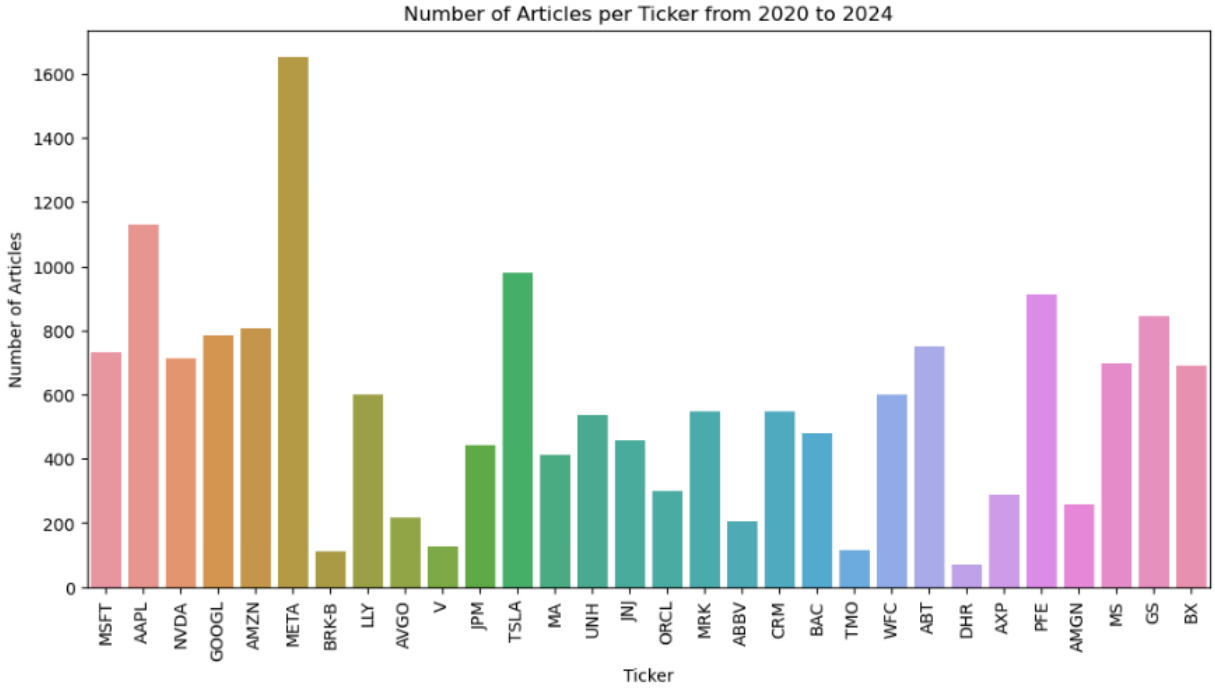


Fig 1.2.1: Distribution of Articles per Company

By performing sentiment analysis on our dataset, using FinBERT we get the following:

Sentiment	*Title*	*Content*
Positive	3737	3538
Neutral	8700	6765
Negative	4581	6715

We follow the same process as in Dataset 1 and we divide the sentiments into the same four categories (Negative, Somewhat Negative, Somewhat Positive, Positive). The distribution for the content and title data points is shown below:

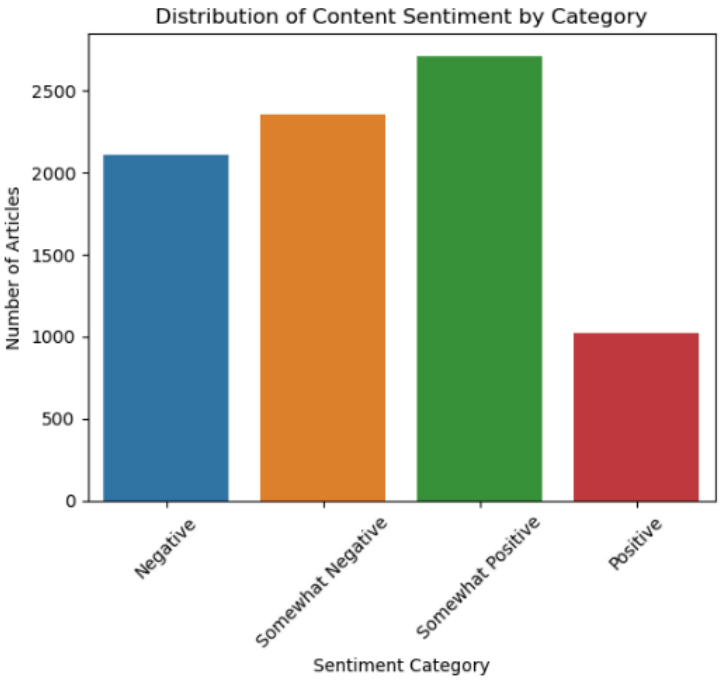


Fig 1.2.2: Number of articles per category for the content sentiments

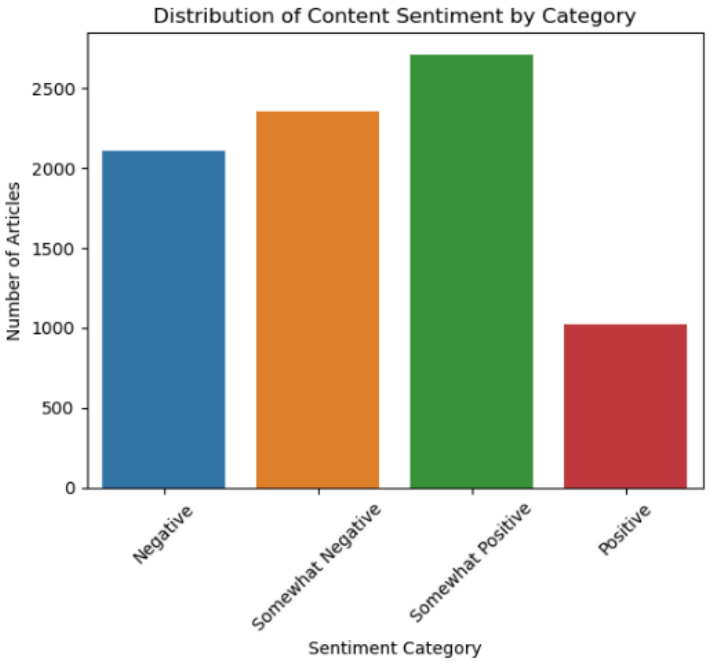


Fig 1.2.3: Number of articles per category for the title sentiments

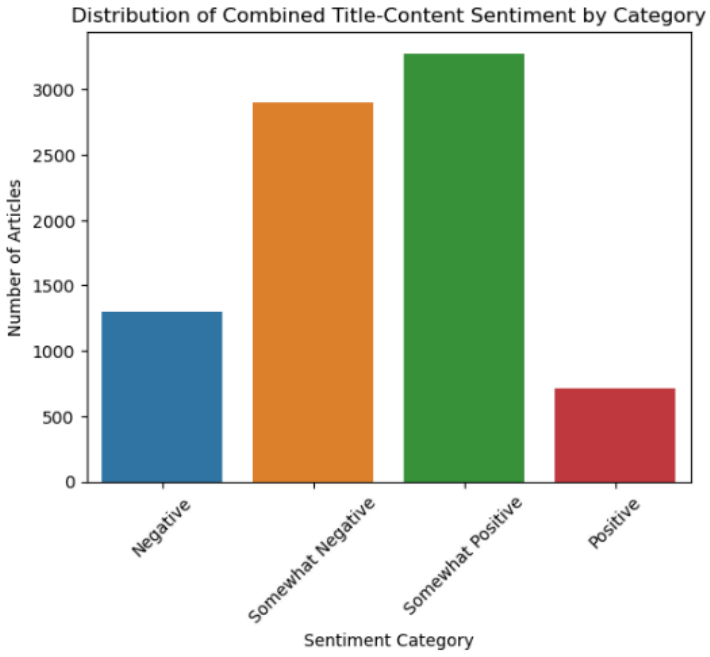


Fig 1.2.4: Number of articles per category for the combined title-content sentiments

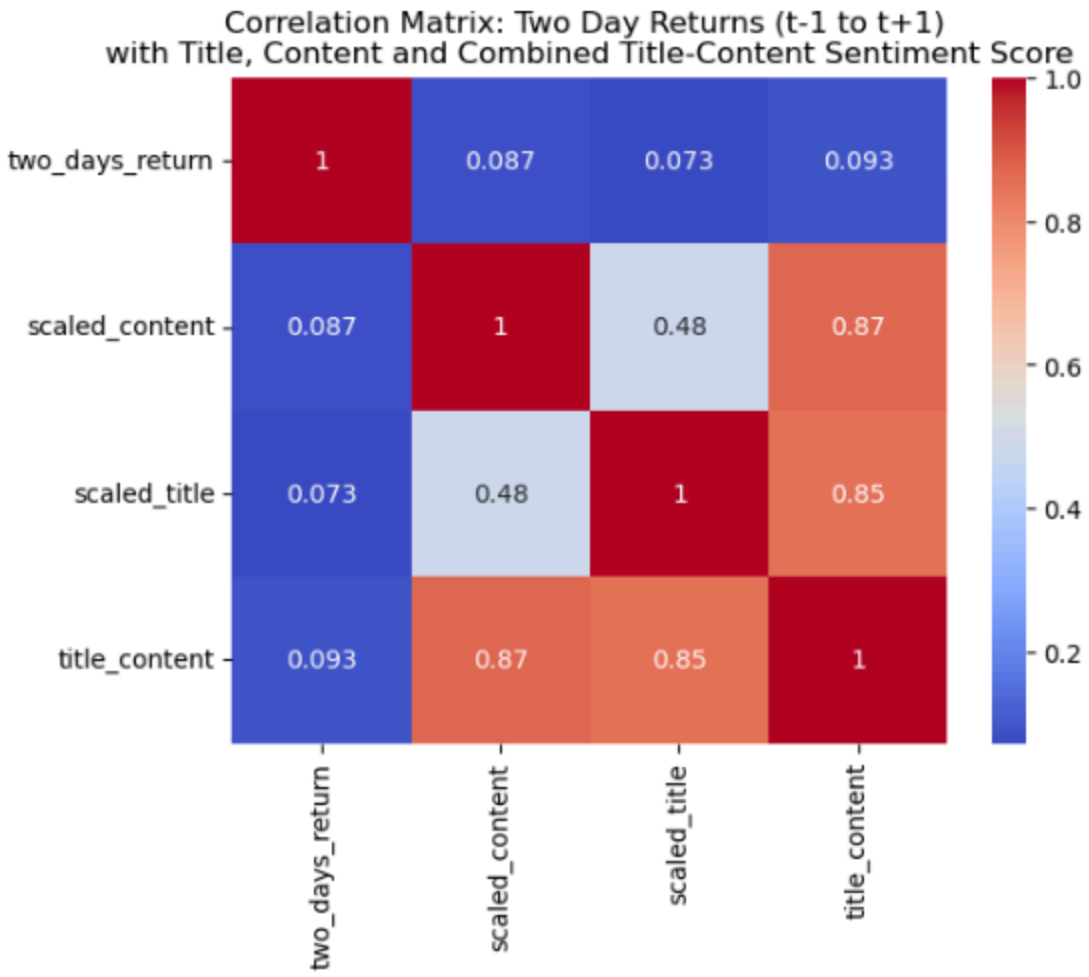


Fig 1.2.5. Correlation Matrix of Two-Day Returns with Title, Content, and the Average of Title and Content Sentiment Scores

Similar to previous dataset, we explore the correlation between news and two-day returns. The correlation matrix also shows the correlation values for the combined title and content sentiment, which is 0.093, is higher than those for both the content and title. On the other hand, when compared to the correlation matrix for the Dataset 1, we can conclude that the sentiment scores derived from Dataset 2 are more correlated to the two-day returns.

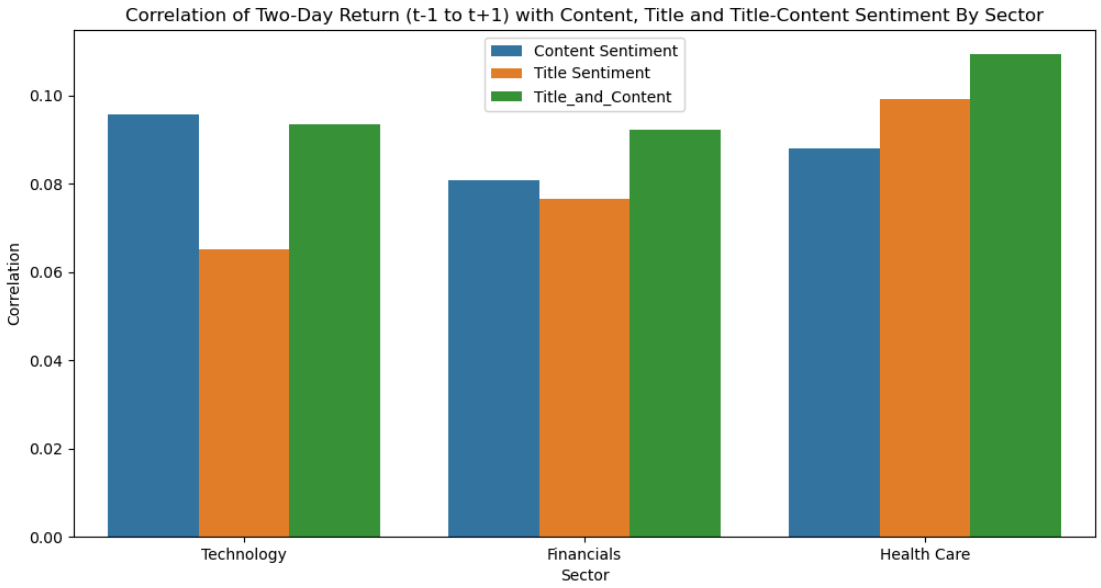


Fig 1.2.6: Correlation of Two-Day Returns (t-1 to t+1) with Content, Title and the Combined Title-Content Sentiment by Sector

When diving into specific sectors in Fig 1.2.6, we found relatively little difference between the correlations for technology and financial sectors, which is very different from that of Dataset 1. This suggests that in articles related to the technology sector, the company mentioned most frequently may not necessarily be the protagonist of the story while the opposite may be true in financial-related articles. Additionally, the sentiments of news titles always present lower correlations for the technology sector and financial sector, which implies that headlines have less influence on two-day return than contents for Dataset 2 in these 2 sectors.

Similar to the previous part, only small sector-based variation exists between sentiment scores and short-term stock returns. These variations can be shown in the following scatterplots:

From Fig 1.2.7, we still cannot see obvious correlations between the content sentiment scores and the two-day return. Similar to that from Dataset 1, the technology sector exhibits the greatest degree of variability among the three sectors, which is also larger than that from Dataset 1. From Fig 1.2.8 and Fig 1.2.9, similar patterns are observed.

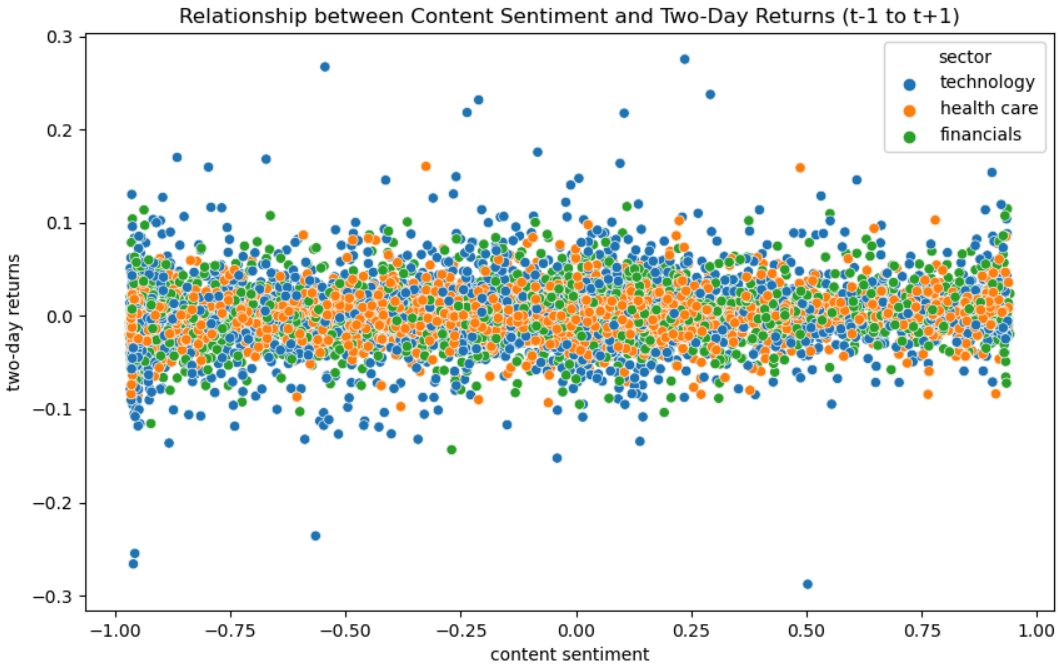


Fig 1.2.7. : Scatterplot of Content Sentiment score vs Two-Day Returns

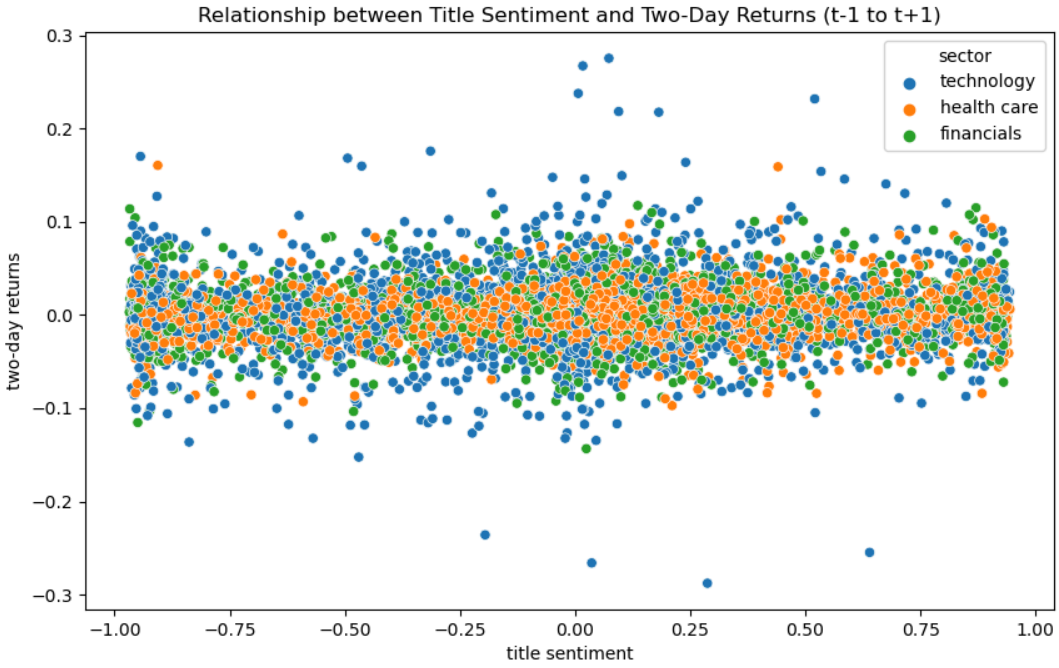


Fig 1.2.8. : Scatterplot of Title Sentiment score vs Two-Day Returns

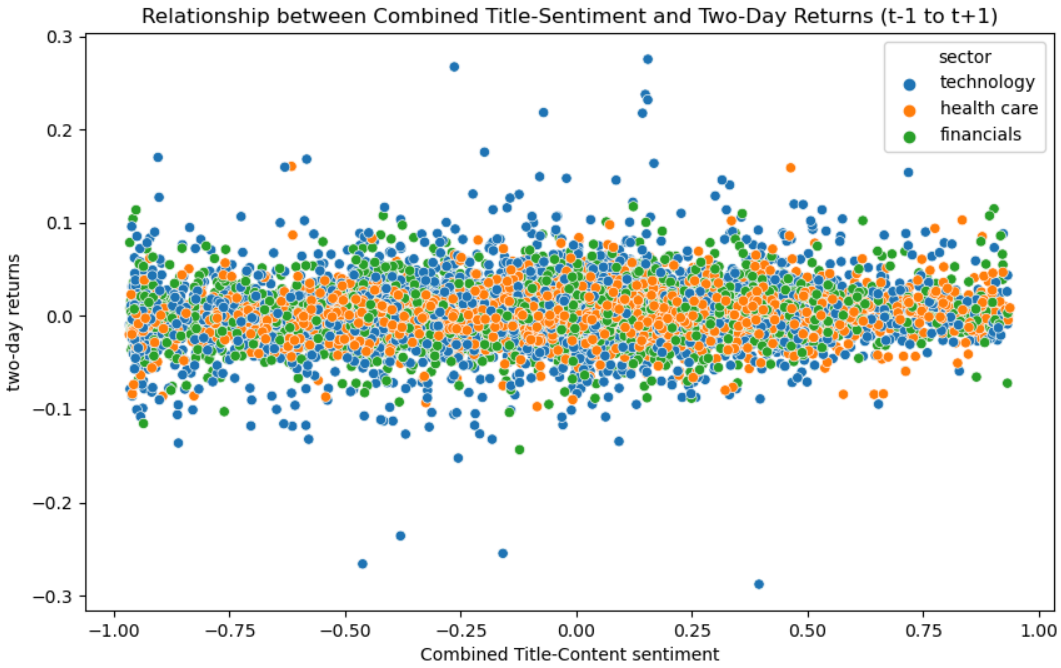


Fig 1.2.9. : Scatterplot of Combined Sentiment score vs Two-Day Returns

By applying a Pearson correlation test we find that the correlation between sentiment and two-day returns, for the content, the title, and the combined title-content sentiment, is statistically significant for the technology and healthcare sectors.

The results are presented in the following table:

		Sector		
		Technology	Financials	Health Care
Content	Correlation	0.096	0.081	0.088
	P-value	2.13e-08	0.000085	0.000015
Title	Correlation	0.065	0.076	0.099
	P-value	0.000137	0.000203	0.000001
Title-Content	Correlation	0.093	0.092	0.109
	P-value	4.55e-08	0.000007	7.28e-08

Table 1.2.1: Pearson Correlation Test: Correlation and p-values for content, title and combined title-content sentiment with two-day returns for each sector

Just like what we conclude from the last section, from the violin plot (Fig 1.2.11), we observe that for each sentiment category corresponding to content sentiments, the majority of the two-day returns are clustered around zero, which aligns with the expectation that daily returns often show small fluctuations. The median return for each category, indicated by the white dot, also appears to be near zero, suggesting no substantial median gain or loss over the two-day period regardless of sentiment.

However, for the "Positive" sentiment category, the range of returns is skewed towards more negative values, which is different from that for Dataset 1. For the "Negative" sentiment category, the stock returns vary in a wide range, which means the market have significantly different responses to obvious negative news. To conclude, the distribution of two-day return performs poorly by content sentiment category.

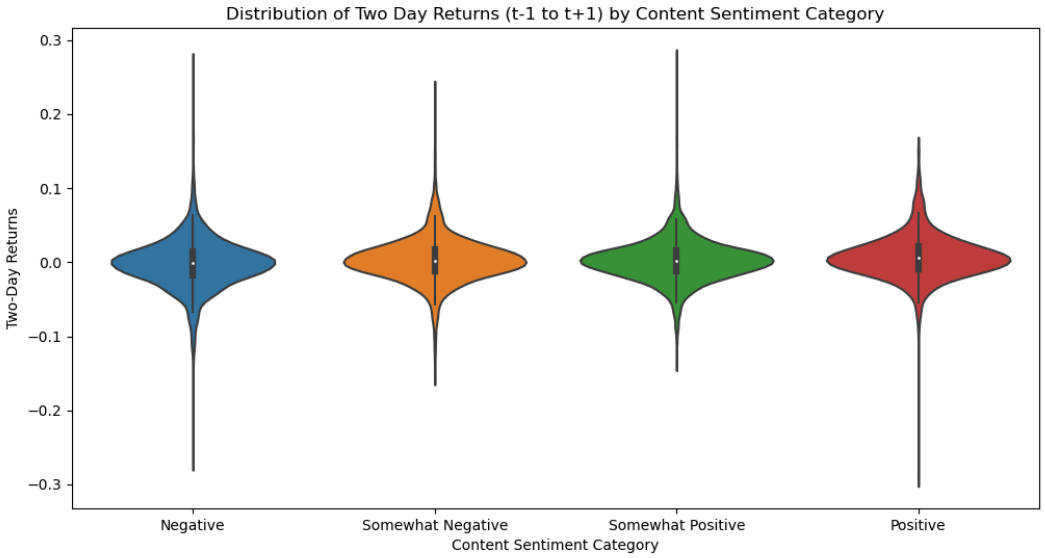


Fig 1.2.11. : Violinplot showing the Distribution of Two-Day Returns by Content Sentiment Category

From Fig 1.2.12, nearly symmetric patterns are observed for four sentiment categories, which suggests that the market has random responses for headlines of articles, no matter where they belong to. While for Fig 1.2.13, for the "Positive" sentiment category, the range of returns is skewed towards more positive values.

For further understanding of the relationships of the sentiment level and two-day returns, we study them in each sector.

Skewed towards more positive returns in 'Somewhat Positive' sentiment category but more negative returns in 'Positive' sentiment category, the technology sector seems to have bigger correlations between content sentiment category and returns from Fig 1.2.14, since

the two-day returns from the other two sectors are likely to be symmetric regardless of where they belong to.

From Fig 1.2.15, patterns are similar except symmetric also appears in the “Somewhat Positive” and “Positive” sentiment category for the technology sector. In another word, the correlations between title sentiment category and two-day return are not as obvious as those of content category.

Skew towards positive returns in “Positive” combined sentiment category happens in all three sectors from Fig 1.2.16, which suggests a greater correlation between combined sentiment category and two-day returns.

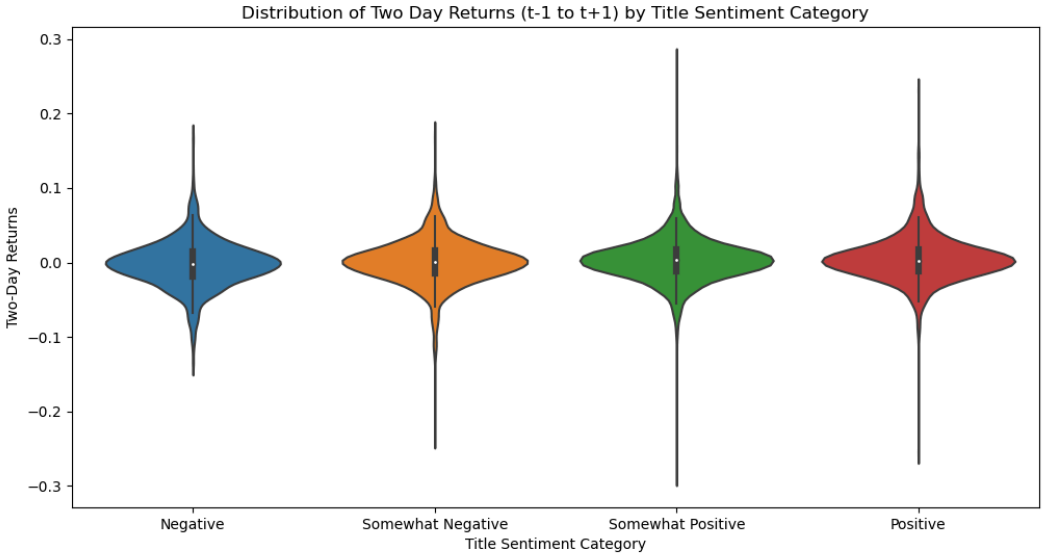


Fig 1.2.12. : Violinplot showing the Distribution of Two-Day Returns by Title Sentiment Category

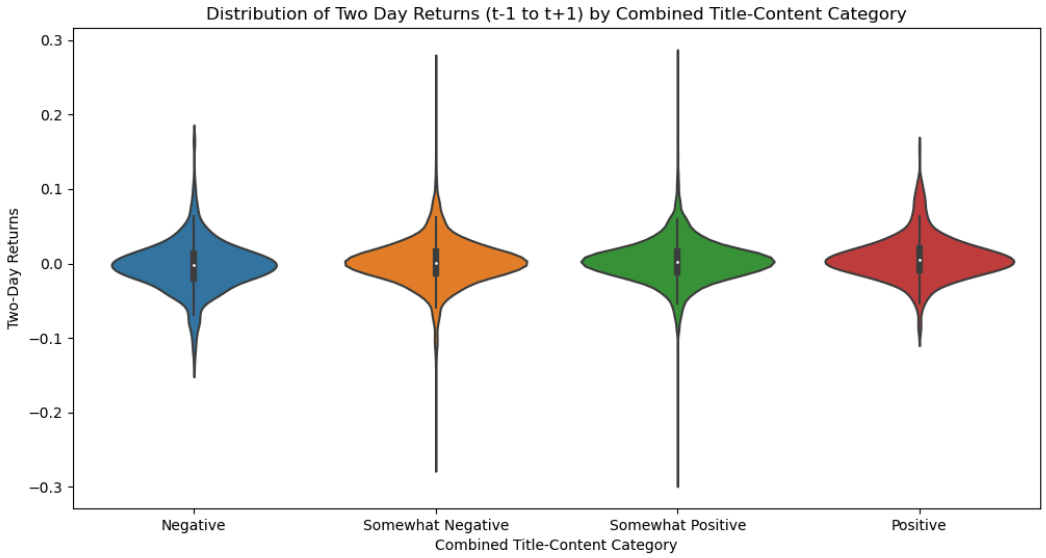


Fig 1.2.13. : Violinplot showing the Distribution of Two-Day Returns by Combined Title-Content Sentiment Category

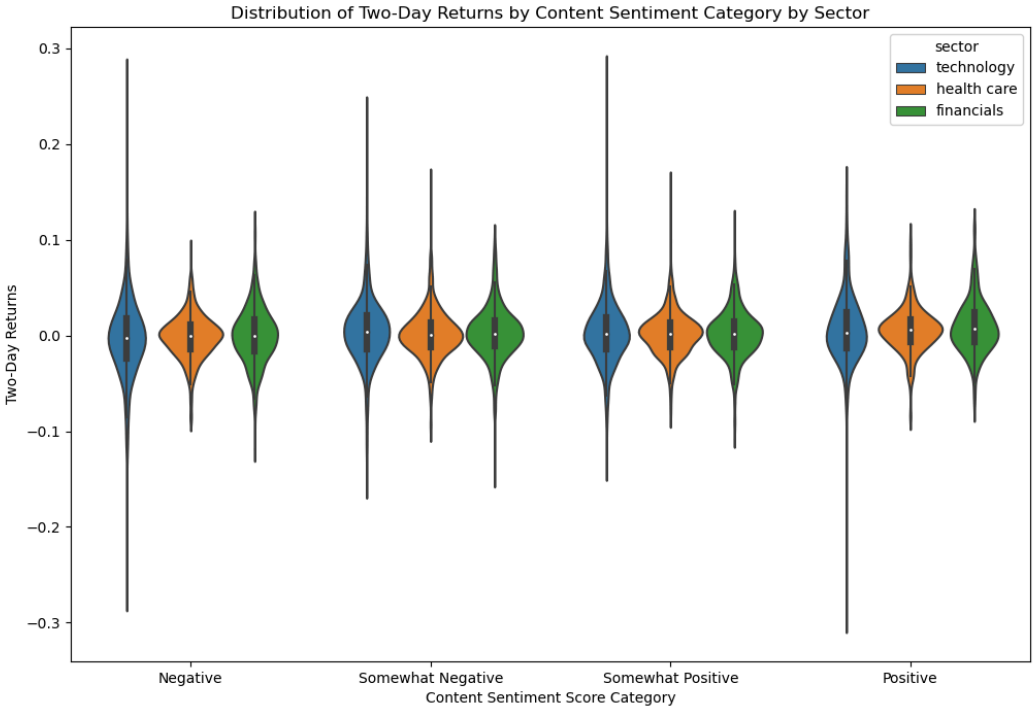


Fig 1.2.14: Distribution of Two-Day Returns by Content Sentiment for each Sector

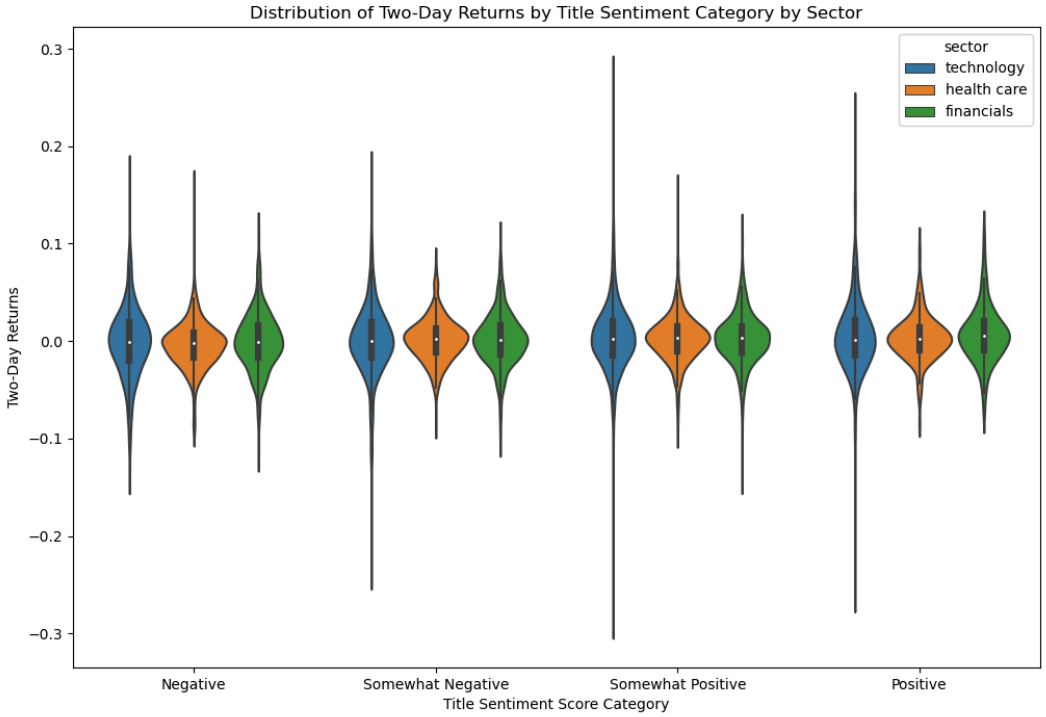


Fig 1.2.15: Distribution of Two-Day Returns by Title Sentiment for each Sector

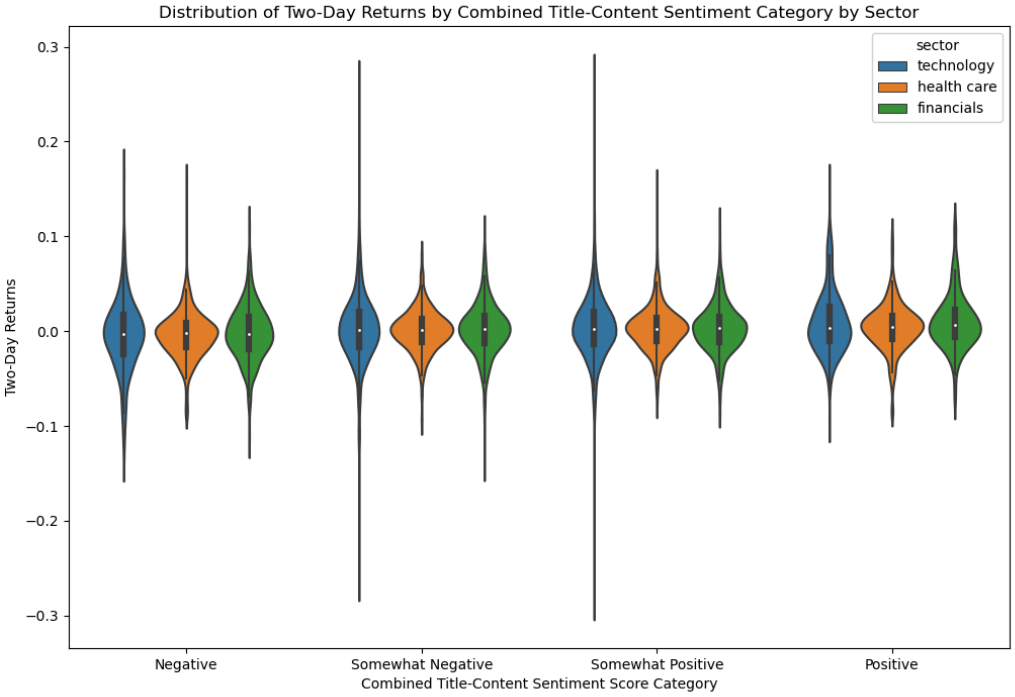


Fig 1.2.16: Distribution of Two-Day Returns by Combined Title-Content Sentiment for each Sector

To facilitate a direct comparison of the average two-day returns across different sentiment categories and sectors for content-related articles the following barplot has been created.

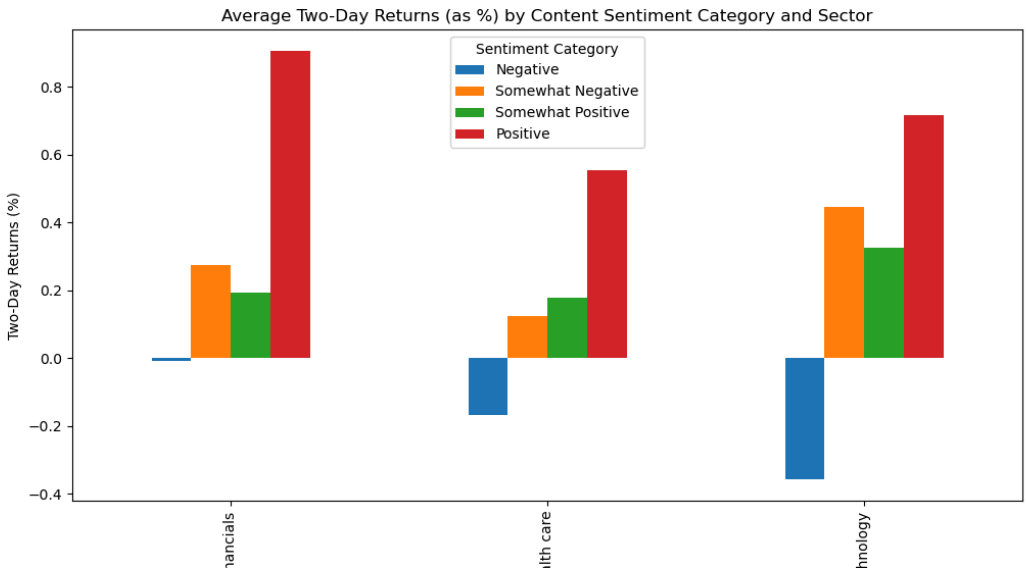


Fig 1.2.17: Average Two-Day Returns (as %) by Content Sentiment for each Sector

The average two-day returns of the ‘Negative’ sentiment category for all sectors is negative and those of other categories are all positive. Among them, the average returns of the ‘Negative’ category for the technology sector is the lowest, while that of the ‘Positive’ category for the financial sector is the highest. Different from what we get from Dataset 1, in Dataset 2, the pattern that “Negative” content sentiment yields negative average returns holds for all of three sectors. However, ‘Somewhat Negative’ content sentiment results in positive returns for all sectors, even has higher average returns than that of ‘Somewhat Positive’ content sentiment for the financial sector and technology sector. This might be explained by the market only responding dramatically and reasonably for news with obvious sentiment tendency(‘Positive’ or ‘Negative’) rather than subtle sentiment(‘Somewhat Positive’ or ‘Somewhat Negative’).

For the sentiments derived from the articles’ titles, the following plot Fig 1.2.18 reveals a similar behavior to that of article content across all sectors for all categories. However, there are some distinctions. The average returns of the ‘Somewhat Negative’ sentiment are all smaller than those of the "Somewhat Positive” sentiment category for all the three sectors. For the health care sector and technology sector, title sentiment still performs well

since ‘Negative’ category has lowest negative average returns in the health care sector and ‘Positive’ category has highest average returns in the financial sector.

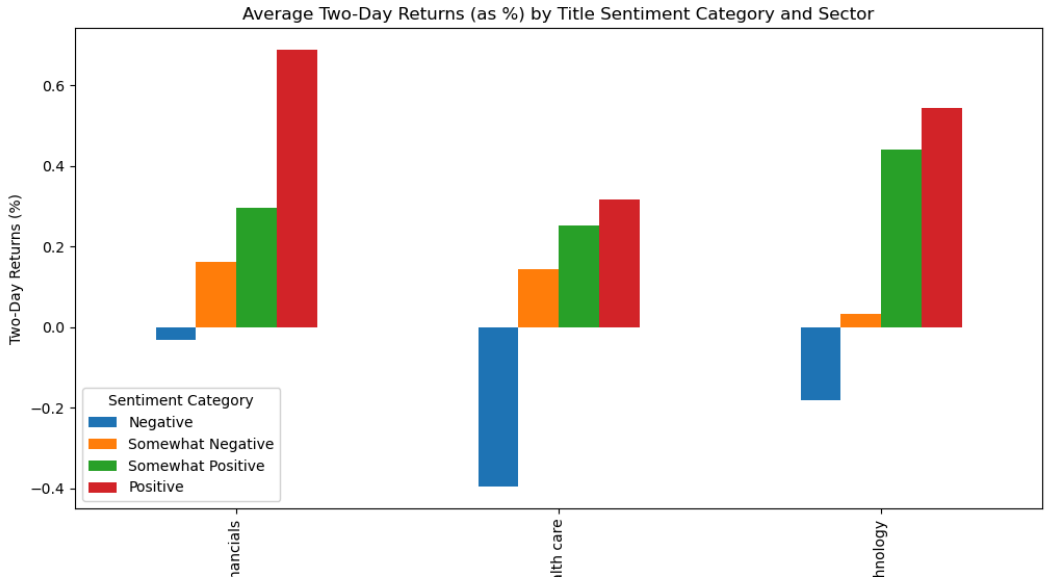


Fig 1.2.18: Average Two-Day Returns (as %) by Title Sentiment for each Sector

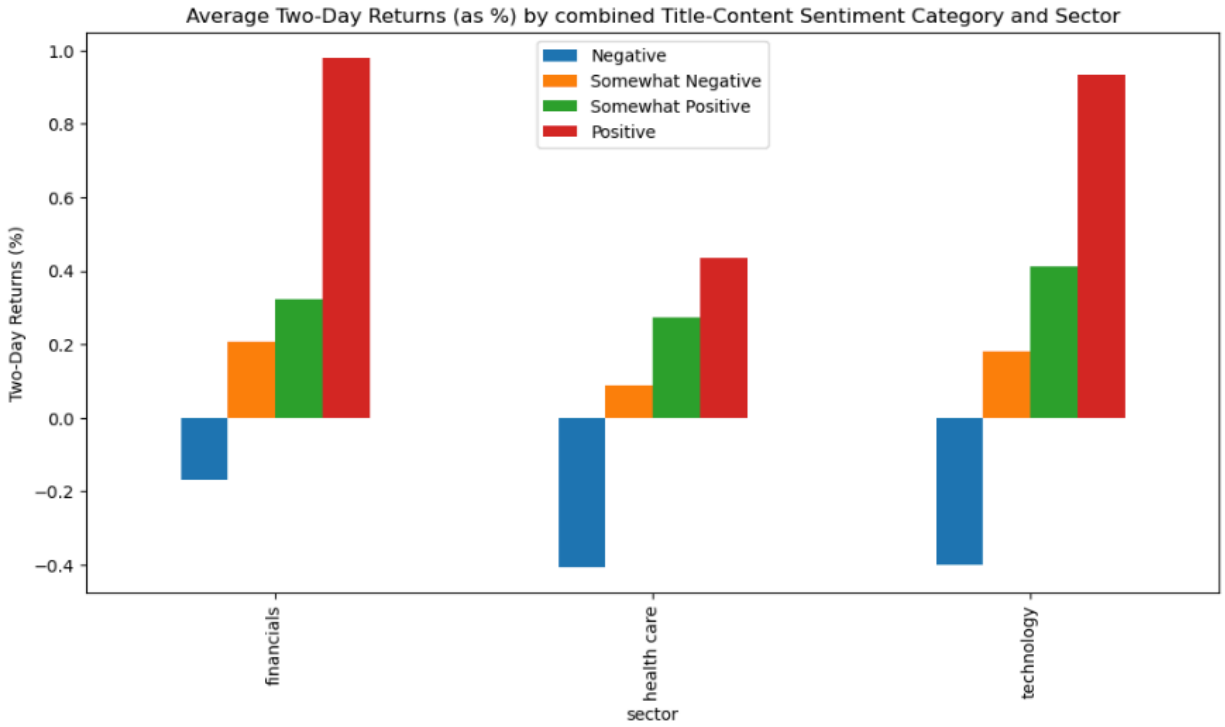


Fig 1.2.19: Average Two-Day Returns (as %) by Combined Title-Content Sentiment for each Sector

The pattern we derived from Fig 1.2.19 is quite similar to that from Fig 1.2.17. The only difference is that the average returns in the health care sector become the lowest , instead of the technology sector.

Conclusions

In this analysis, we investigated the correlation between news sentiment and two-day stock returns, comparing two datasets. We found that the combined sentiment of news titles and content exhibits a higher correlation (0.093) with two-day returns compared to individual title or content sentiment. Dataset 2 showed stronger correlations between sentiment scores and stock returns than Dataset 1.

When examining specific sectors, we observed relatively consistent correlations across technology and financial sectors, indicating that frequently mentioned companies in technology-related articles may not necessarily be the story protagonists, unlike in financial articles. However, title sentiments generally showed lower correlations with two-day returns in technology and financial sectors, suggesting that headlines have less influence on returns compared to content in these sectors.

Although small variations existed between sector-based sentiment scores and returns, the technology sector displayed the greatest variability, particularly in Dataset 2. Statistical tests confirmed significant correlations between sentiment and returns in the technology and healthcare sectors.

Analysis of sentiment category distributions revealed that while the majority of two-day returns clustered around zero, positive sentiment categories exhibited a skew towards more negative returns, especially in Dataset 2. Furthermore, the healthcare sector demonstrated stronger correlations between content sentiment and returns, particularly in positive sentiment categories.

In conclusion, sentiment analysis provides valuable insights into stock returns, with combined title-content sentiment showing the strongest correlation. However, variations exist between sectors, emphasizing the need for sector-specific analysis in understanding sentiment-return relationships.

1.2.3. Dataset 3: Multiple companies per article

Dataset 3 consists of 14283 data points where we weight each ticker by the importance of each article, calculated as the percentage of mentions of each company divided by the total number of mentions

The distribution of these articles across different tickers is visualized in the following barplot. The distribution of articles per company, ordered by their market capitalization from largest to smallest, is shown below.

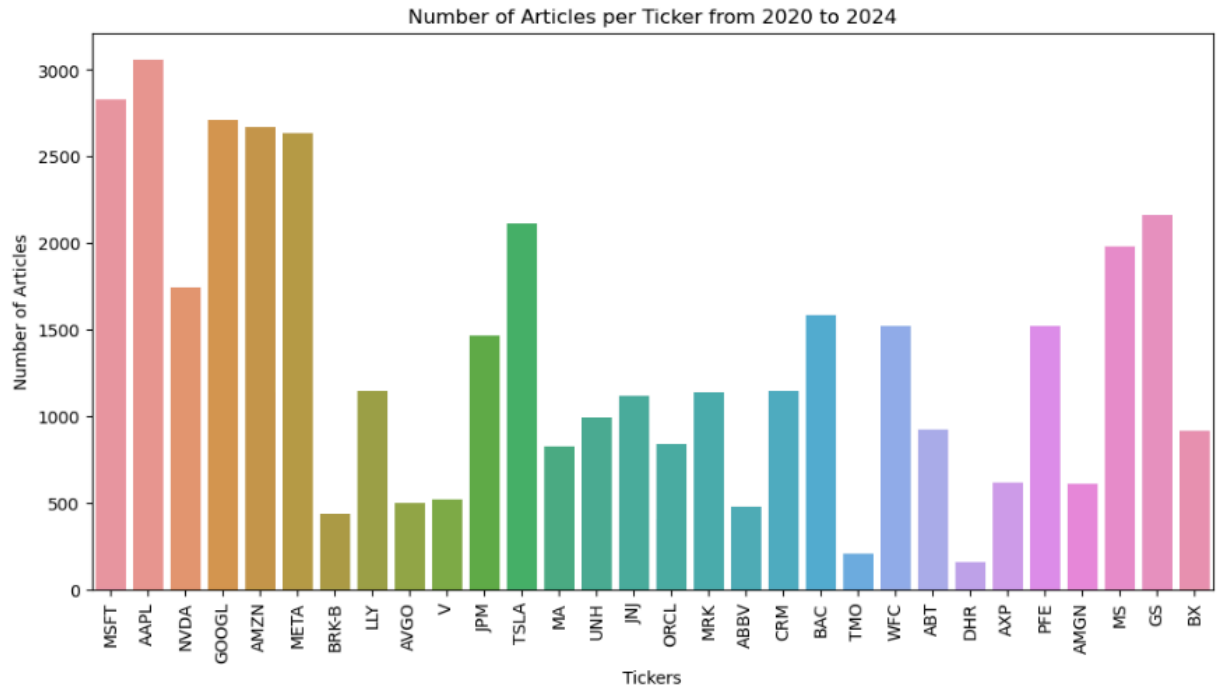


Fig 1.3.1: Number of articles per stock's ticker ordered by companies' market cap

By performing sentiment analysis on our dataset, using FinBERT we get the following:

Sentiment	*Title*	*Content*
Positive	*8784*	*9342*
Neutral	*20925*	*14481*
Negative	*10878*	*16764*

First, we can look at the visualization of the correlation matrix and we can see that there is a slightly higher correlation between the two day return with the content then there is with the title, even though the scores are very similar. We can also see that there is a significant

correlation between the sentiment of the content and the sentiment of the title, which makes sense considering that the sentiment should be for the same article.

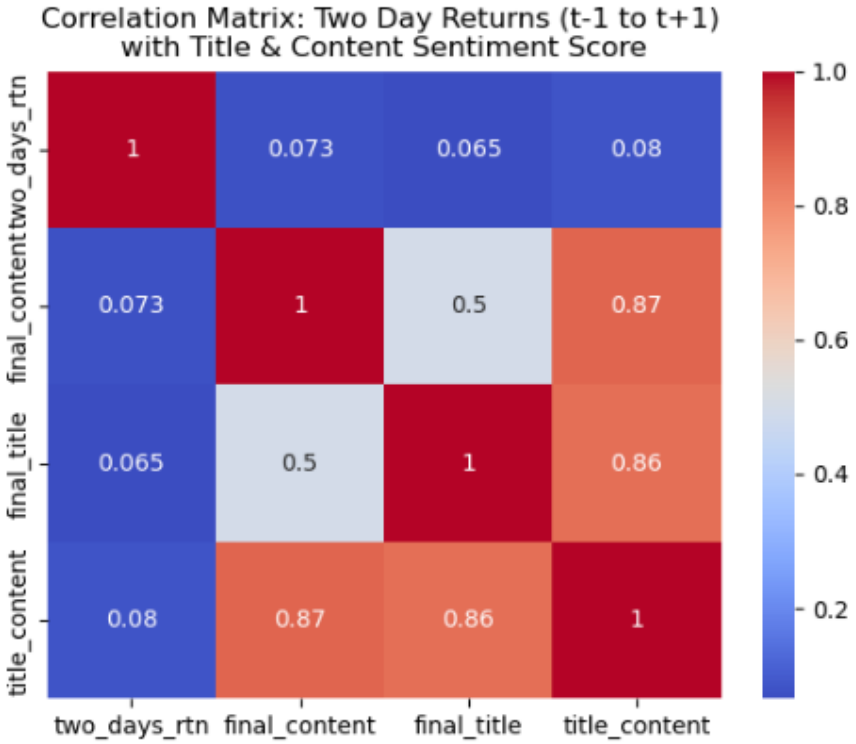


Fig 1.3.2. Correlation Matrix of Two-Day Returns with Title, Content, and Average of Sentiment Scores

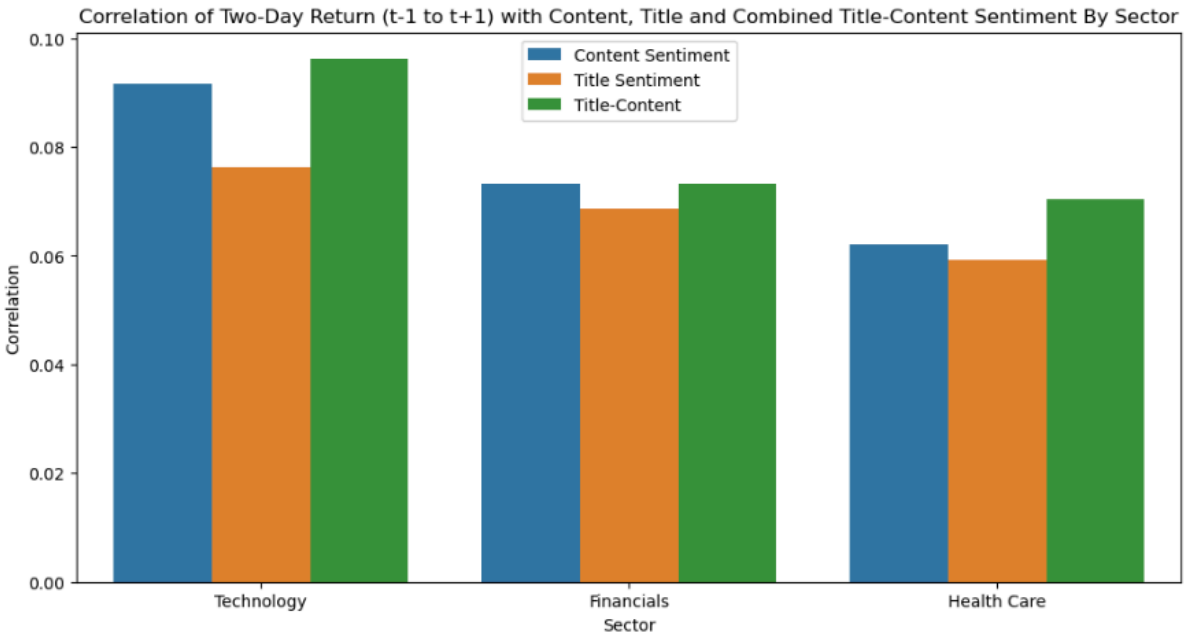
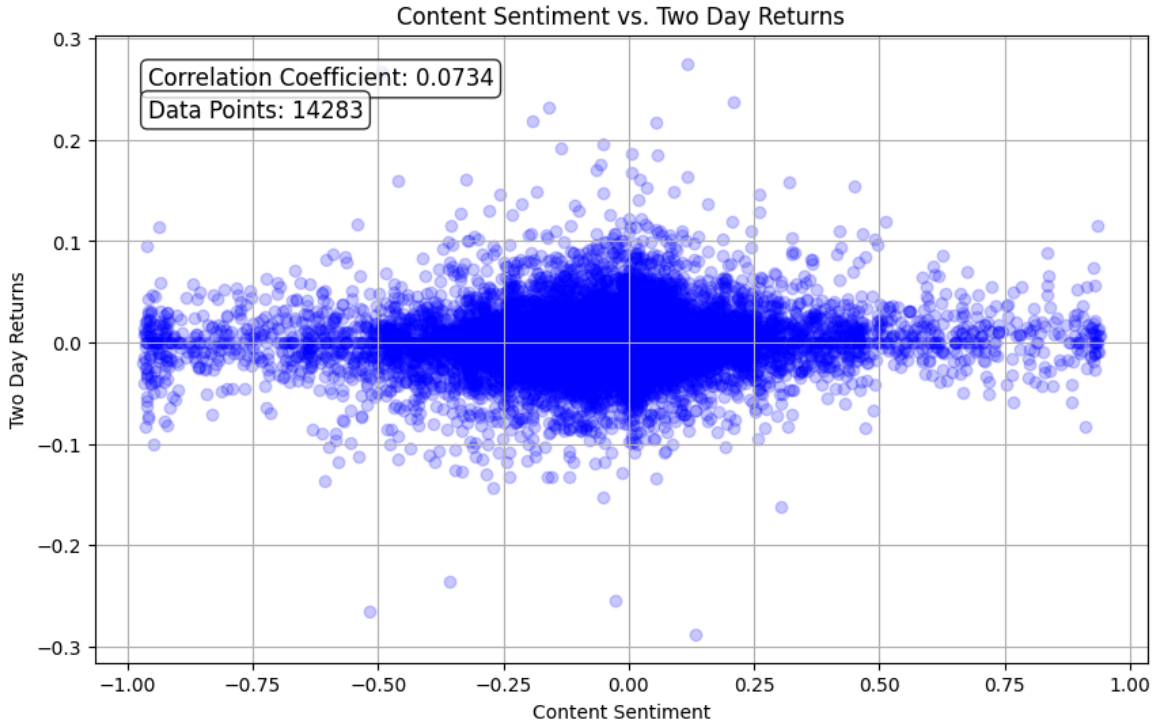
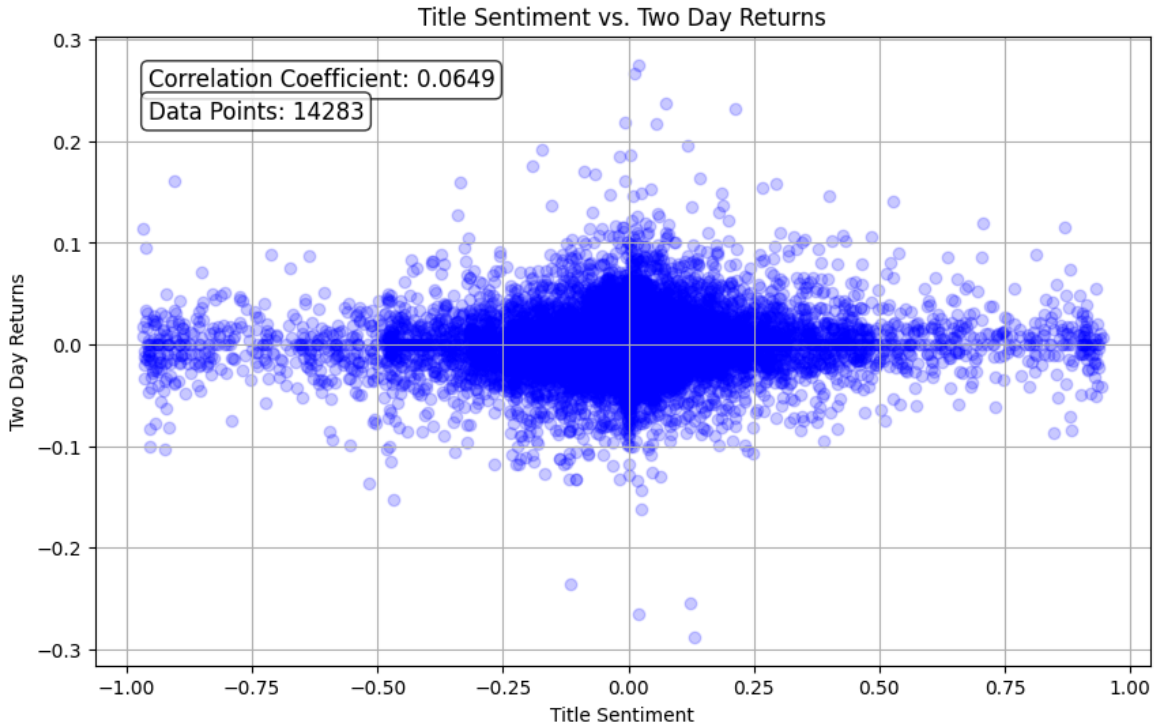


Fig 1.3.3: Correlation of Two-Day Returns (t-1 to t+1) with Content, Title and the Combined Title-Content Sentiment by Sector

The correlation of the sentiment and the two day returns for the content and for the title is shown below.

Figure 1.3.4/5 Scatter Plot for Title and Content of all articles and tickers vs two day returns.





Next we split the data for the content sentiment by industry to try to find an underlying relationship in the data.

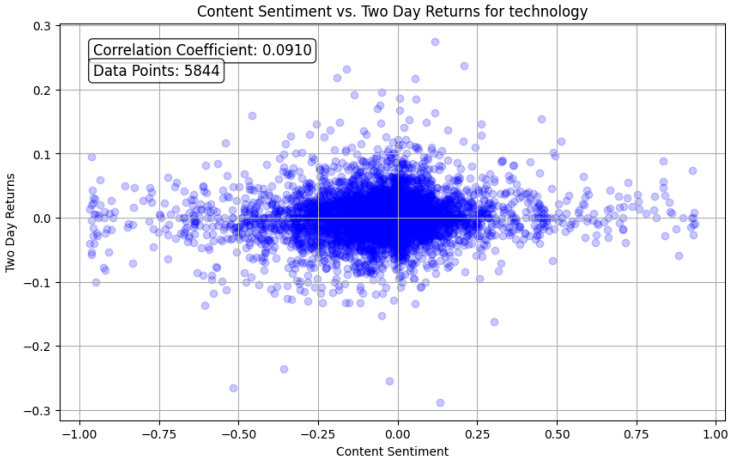


Figure 1.3.6 Content sentiment vs two day returns for technology tickers only.

This sector had a noticeably higher correlation compared Fig 1.3.4, which may make sense due to the volatile nature of technology stock prices.

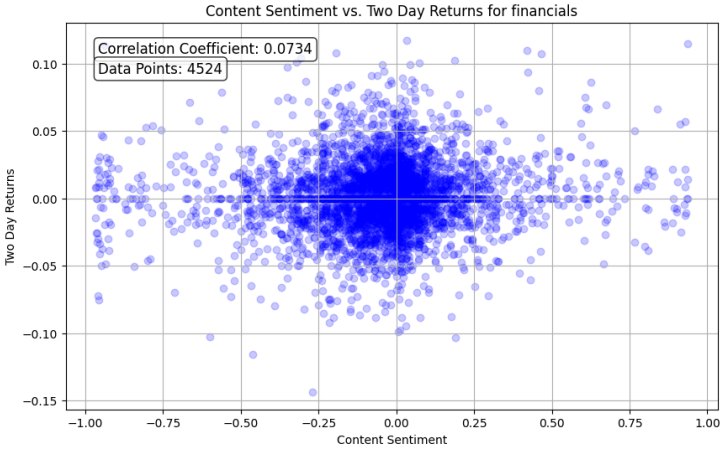


Figure 1.3.7 Content sentiment vs two day returns for financial tickers only.

This sector had about the same correlation compared to Fig 1.3.4, which was expected since FinBert, the model we used for the study, is trained on financial data.

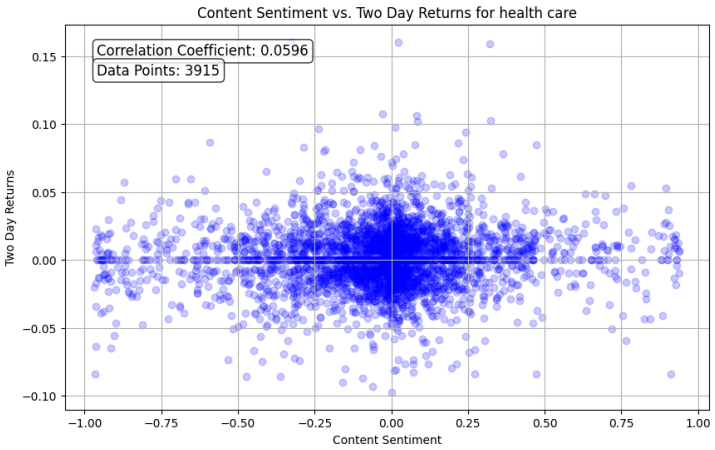


Figure 1.3.8 Content sentiment vs two day returns for health care tickers only.

This sector had about significantly lower correlation compared to Fig 1.3.4, which was unexpected and accounts for a significant pattern that health care returns are much less correlated compared to all other stocks.

Next we split the data for the content sentiment by industry to try to find an underlying relationship in the data.

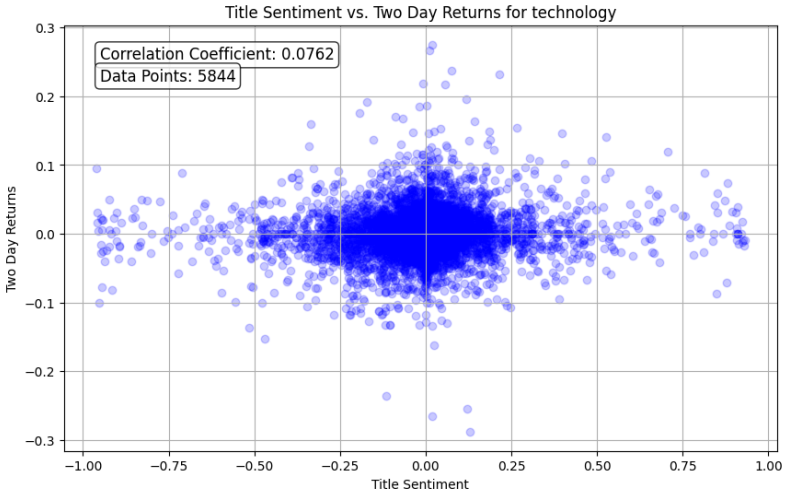


Figure 1.3.9 Title sentiment vs two day returns for technology tickers only.

This sector had a noticeably higher correlation compared to Fig 1.3.5, which may make sense due to the volatile nature of technology stock prices and lines up with what we expected from Fig 1.3.6.

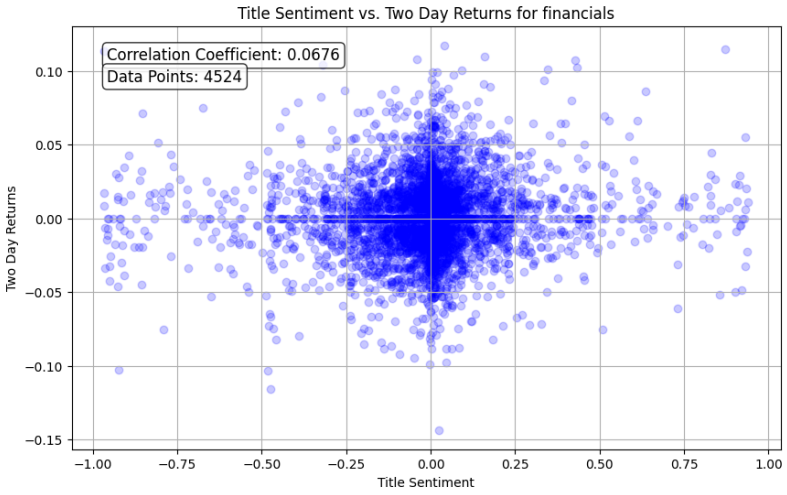


Figure 1.3.10 Title sentiment vs two day returns for financial tickers only.

This sector has a similar correlation to the correlation in Fig 1.3.5 and 1.3.7, which is expected due to the nature of using the FinBert model.

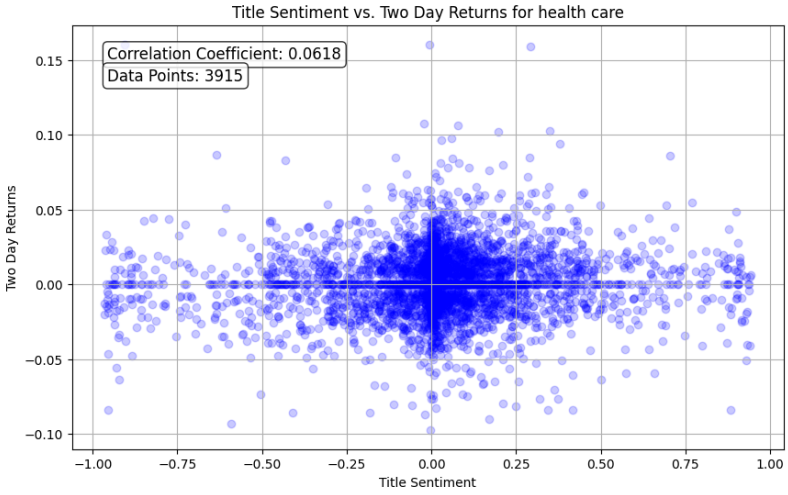


Figure 1.3.11 Title sentiment vs two day returns for health care tickers only.

This sector has a similar correlation to the correlation in Fig 1.3.5, both of which are significantly lower then Fig 1.3.8. This stands out as a notable outlier as a sector for the dataset as a whole.

Next we will use violin plots to see the distributions of returns to see the difference in variance, mean, and quartiles to better see the spread of two day returns and sentiment.

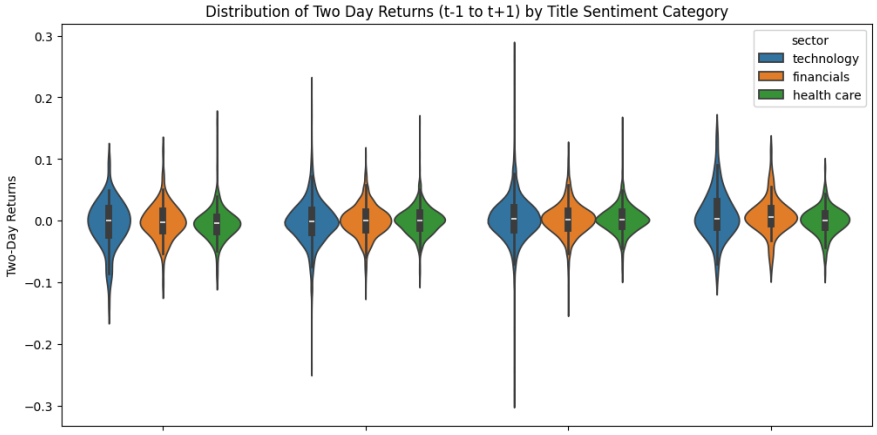


Fig 1.3.12 Distribution of two day returns by title sentiment and sector.

We can see in the graph that the most variance appears in the tech sector.

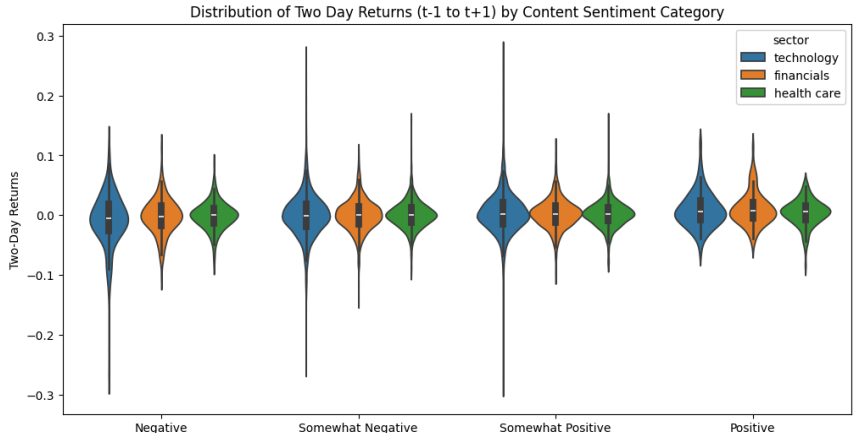


Fig 1.3.13 Distribution of two day returns by content sentiment and sector.

Another interesting point that stood out is that negative articles were more extreme when considering content vs title.

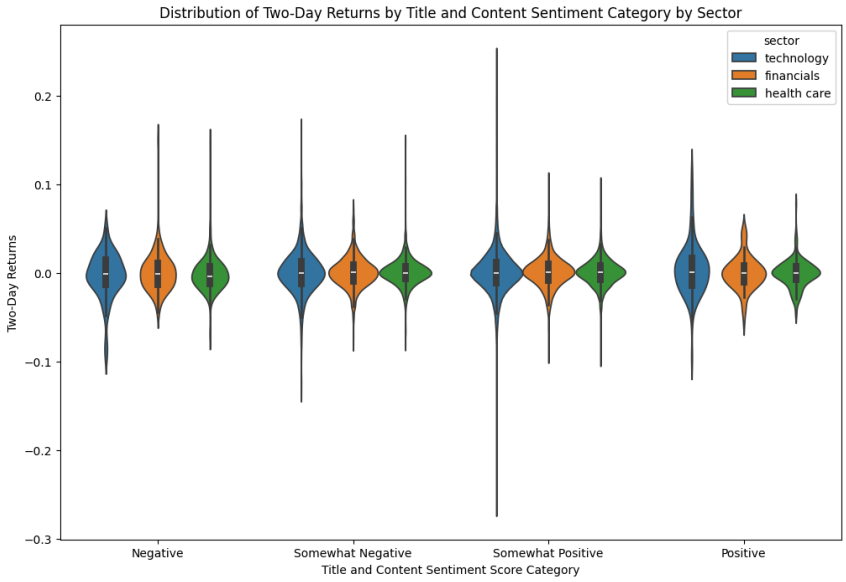


Fig 1.3.14 Distribution of two day returns by content and title sentiment and sector.

And here we have the distribution for both the title and content as well.

To better visualize the most extreme cases of the dataset, we looked at the data now through a scatterplot by sector, and we saw an interesting pattern with the extremes.

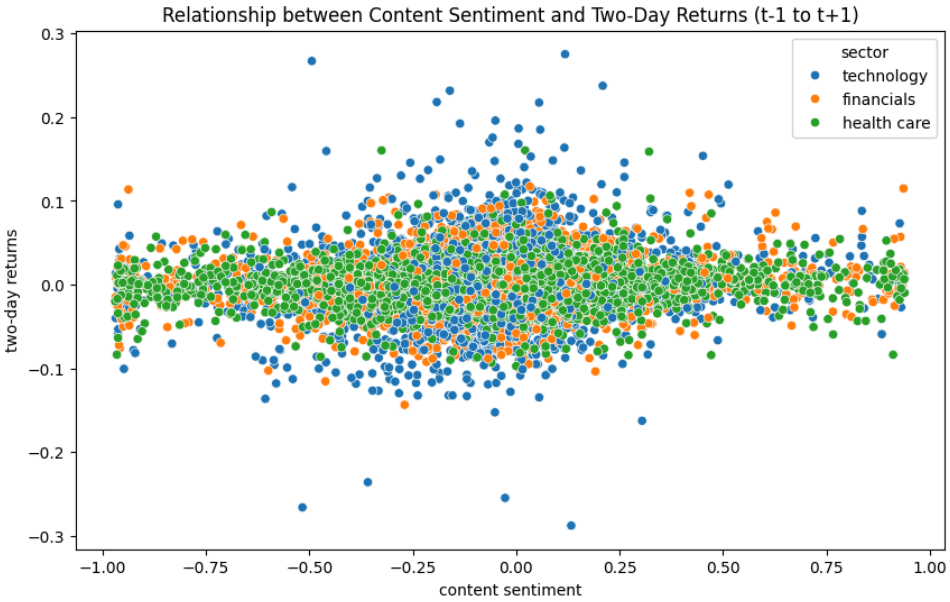


Fig 1.3.14 Relationship between content sentiment and two day returns.

As we can see, the majority of outliers are from the technology sector.

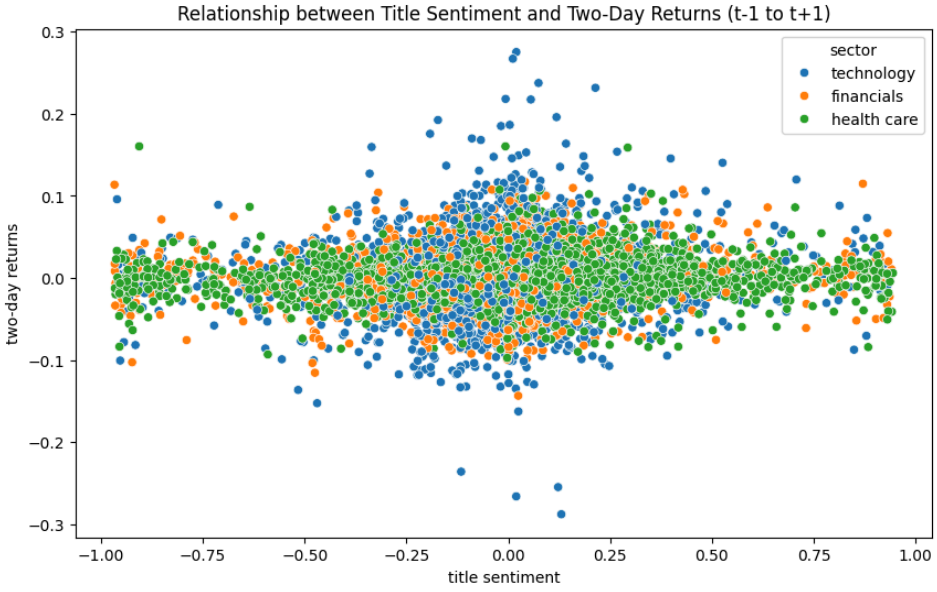


Fig 1.3.15 Relationship between title sentiment and two day returns.

This pattern continues in the title sentiment as well, showing the difference between technology returns and other sectors.

Below is a representation of the scatterplot with the combined title/content sentiment for two day returns, where we see another interesting trend where the technology sector has the most extreme points in the dataset, something we observed slightly earlier, but stands out a lot more in this graph. While this makes sense with the volatility of the tech field, it better helps explain the difference in correlation for the tech sector relative to other sectors.

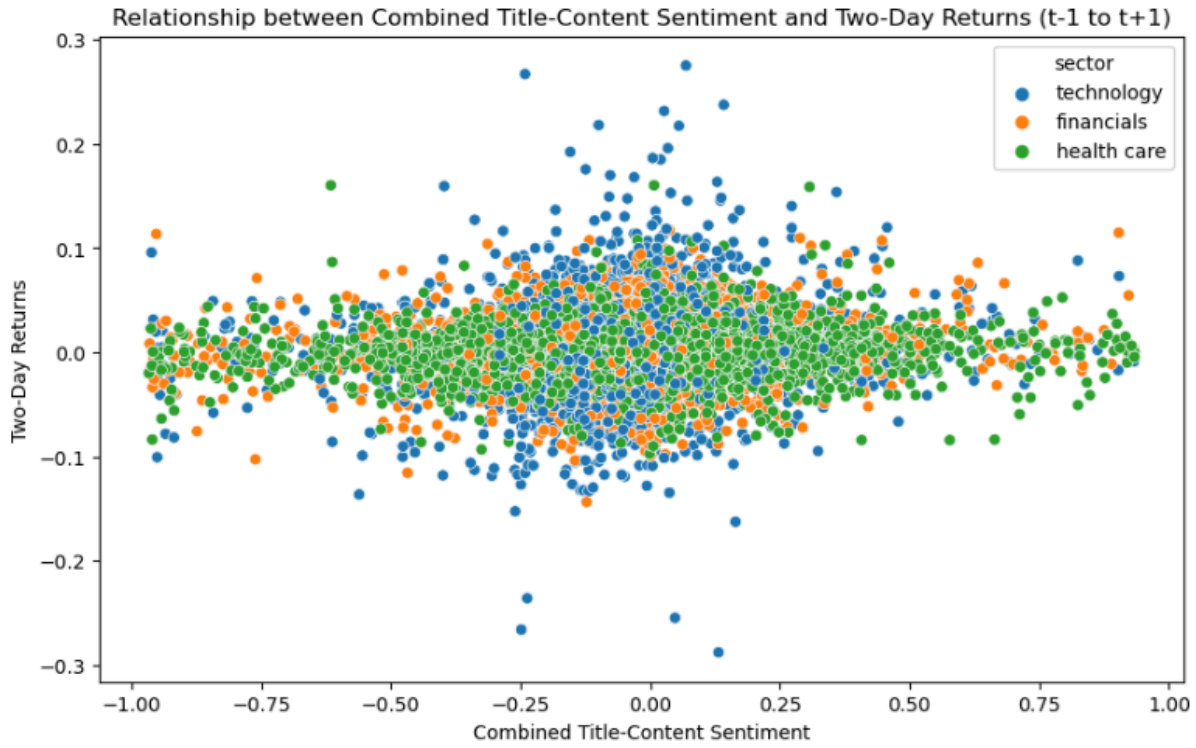
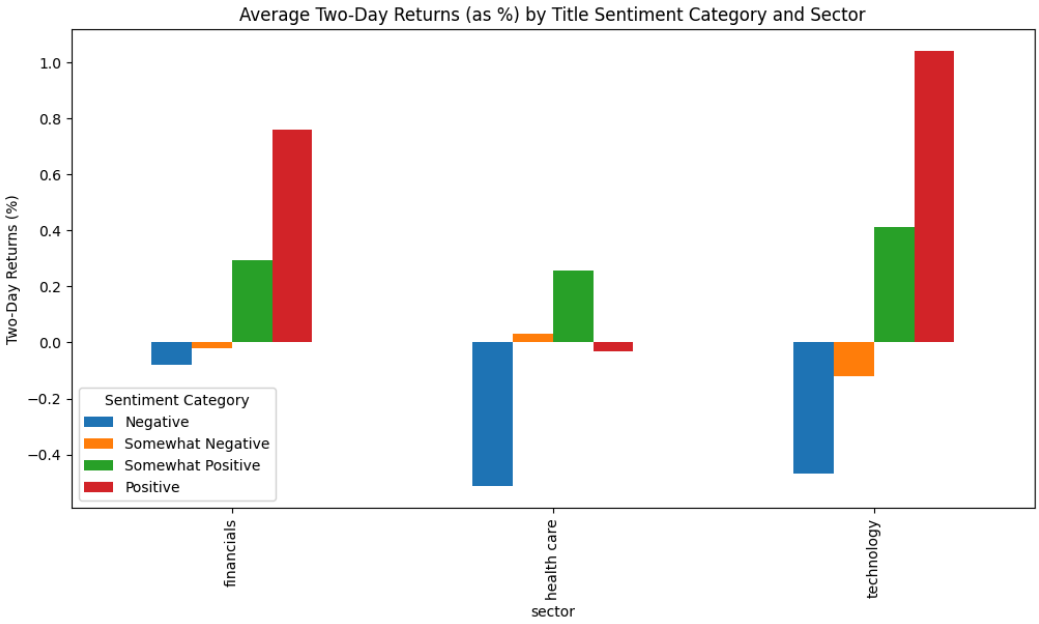
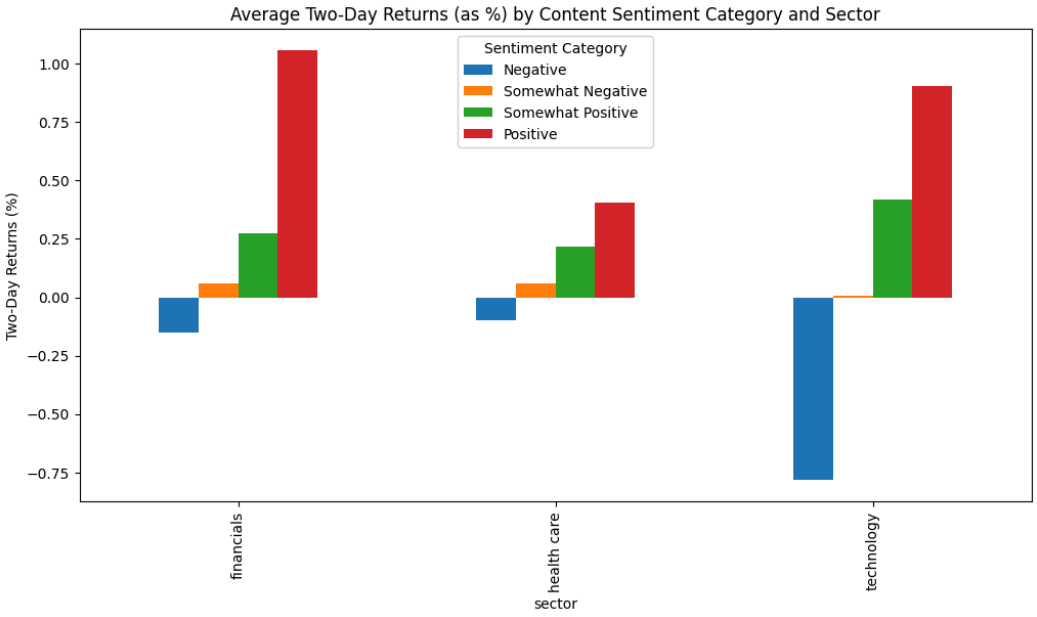


Fig 1.3.16 Relationship between the combined title and content sentiment and two day returns

Then, when we sort the data by sentiment polarity and sector, we get some interesting results. The first thing that stands out is that the healthcare sector has the smallest magnitude of change in returns, regardless of article sentiment, which explains why the correlation is so low. The second notable thing is that the tech industry has the most polarizing articles, which helps explain Fig 1.3.6 and Fig 1.3.9. Moreover, we can notice that somewhat positive and somewhat negative articles both positively impact returns, while only fully negative articles negatively impact returns and somewhat negative articles have almost no impact, which is very interesting as well.

Fig 1.3.17, Fig 1.3.18, Fig 1.3.19 Average two day returns by title and content sentiment by category and sector.





Finally, the table shows the summary of the values found for Dataset 3 where every ticker is included regardless of how many times it's mentioned in the article.

Table 1.3.1 Relationship between Content and Title Correlation and P value by Sector.

		Sector		
		Technology	Financials	Health Care
Content	Correlation	0.09159	0.073253	0.05924
	P-value	2.30e-12	8.12e-07	0.000209
Title	Correlation	0.07631	0.06855	0.06210
	P-value	5.19e-09	0.000004	0.000101
Title-Content	Correlation	0.096	0.0825	0.070
	P-value	1.67e-13	2.68e-08	0.0001

1.2.4. Conclusions

In the following tables, we present the overall results from the analysis of three different datasets.

First, Table 1.4.1. shows the total correlation between the two-day returns and the sentiments for content, title, and their combination for Datasets 1,2 and 3. For all categories, highlighted in green, the highest correlation was observed in Dataset 2, which includes articles where sentiment scores are aligned with the most frequently mentioned company. The correlations for this dataset are 0.087 for the content and slightly lower at 0.073 for the titles, with the highest correlation observed for the average sentiment of the title and content. Overall, these values indicate a weak but significant association between news sentiment and financial returns.

Category	Dataset	Correlation
Content	1	0.077
	2	0.087
	3	0.073
Title	1	0.071
	2	0.073
	3	0.065
Title-Content	1	0.085
	2	0.093
	3	0.08

Table 1.4.1 Overall Correlation for Datasets 1,2,3 for Content,Title and Title-Content Average

Next, we present the overall results for each sector. From our analysis of the content of the articles, the technology sector demonstrates a higher correlation between two-day returns and news sentiment across all categories. Specifically, Dataset 1 - where each article mentions one company - shows the highest correlation at 0.106. Additionally, for title sentiments and two-day returns, the health care sector shows a stronger correlation in Datasets 1 and 2. In contrast, for Dataset 3 - which includes multiple companies weighted by mention frequency - the technology sector again exhibits a higher correlation. Similarly, for the combined category of title and content average, we observe consistent trends. In the first two datasets the health sector appears more correlated, while in the third dataset, the technology sector shows the highest correlation with a value of 0.096.

Category	Dataset	Values	Sector		
			Technology	Financials	Health Care
Content	1	Correlation	0.106	0.046	0.075
		P-value	0.000057	0.116	0.0043
	2	Correlation	0.096	0.081	0.088
		P-value	<0.00001	0.000085	0.000015
	3	Correlation	0.092	0.073	0.059
		P-value	<0.00001	<0.00001	0.00021
Title	1	Correlation	0.067	0.039	0.117
		P-value	0.011	0.177	0.000008
	2	Correlation	0.065	0.076	0.099
		P-value	0.00014	0.0002	<0.00001
	3	Correlation	0.076	0.069	0.062
		P-value	<0.00001	<0.00001	0.0001

Title -Content	1	Correlation	0.097	0.049	0.11
		P-value	0.0002	0.087	0.000022
	2	Correlation	0.093	0.092	0.109
		P-value	<0.00001	<0.00001	<0.00001
	3	Correlation	0.096	0.082	0.07
		P-value	<0.00001	<0.00001	0.00001

Table 1.4.2: Correlation and p-value for content and title sentiments with two-days returns for each sector and each dataset.

We continue our analysis by exploring the different topics that our articles present by employing Topic Modeling.

1.3. Topic Modeling

Topic Modeling is an unsupervised machine learning technique for finding abstract topics within a collection of documents. The model uses algorithms to identify clusters of similar words within a text and categorize them based on similarity. The most frequent topic modeling methods are LSA (Latent Semantic Analysis), LDA (Latent Dirichlet Allocation), Top2Vec and LLM model BERTopic. In this project we implement BERTopic for out topic modeling.

BERTopic is the most recent state-of-the-art topic modeling technique that utilizes transformers to identify topics in large text documents. It leverages BERT and c-TF-IDF to create clusters that represent and interpret the different topics while retaining significant words in topic descriptions. In this project we decided to use this model compared to others as it provides some advantages.

1. Interpretable Topics: BERTopic has been shown to produce more interpretable topics compared to other models like LDA or Top2Vec.

2. Captures semantic relationships between words: BERTopic utilizes transformers' ability to capture the semantic relationship between words, which results in more accurate and meaningful topics. That means that BERTopic is designed to understand the meaning and contextual relationship between words rather than relying solely on the frequency of each word. One key difference between BERTopic and LDA is that the former features continuous topic modeling whereas LDA provides discrete modeling.

3. Minimal pre-processing requirements: although preprocessing of text data is always a good practice before feeding data to any model, using an LLM requires the minimum effort of pre-processing techniques, as these models are already pre-trained to handle this part. Other models' reliability relies on our capability to clean and pre-process the data, using tokenization, lemmatization, removal of stop words etc. where in most cases this is a pivotal challenge with unstructured data. BERTopic leverages advanced topic modeling techniques that efficiently manage these preprocessing steps.

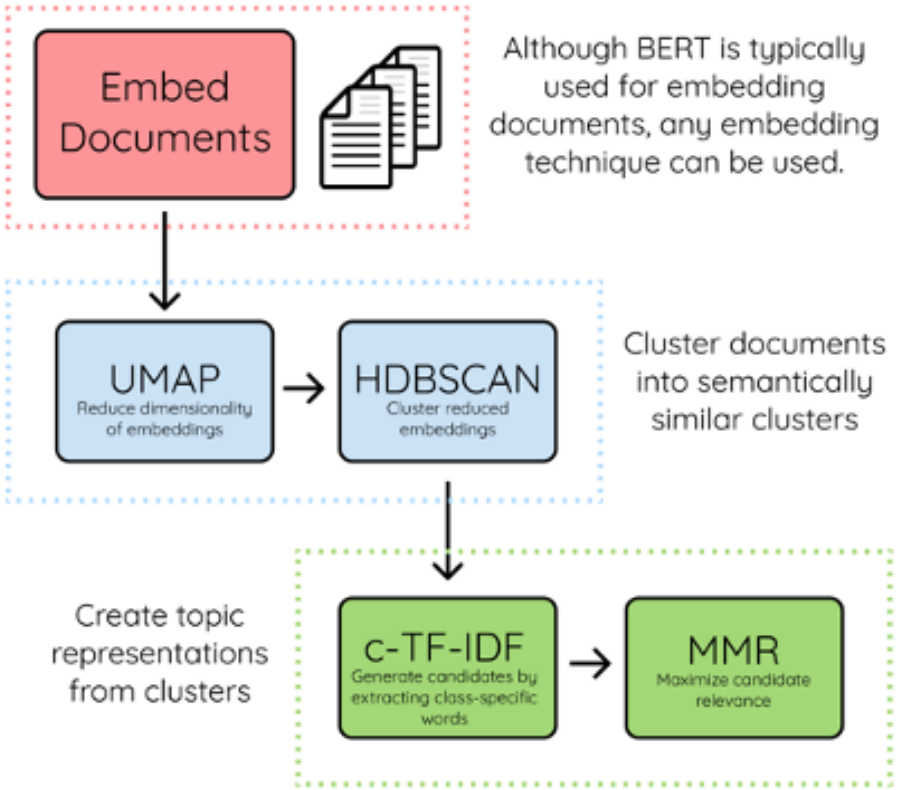
4. Fine-grained control of the number of topics: Unlike other models, BERTopic enables users to precisely adjust the quantity of topics extracted, offering detailed customization that is especially valuable in situations where determining the exact number of topics is essential. Although BERTopic doesn't require a specification of the number of topics, it offers a hierarchical reduction mechanism to merge topics based on similarities, adjusted for research needs.

5. Computational Efficient and Fast for long documents while LDA is faster for a small number of topics

How BERTopic works

The BERTopic model follows five steps:

1. Embeddings' extraction
2. Dimensionality Reduction using UMAP
3. Cluster embeddings reduction using HDBSCAN
4. Topic words extraction using c-TF-IDF
5. Maximum Marginal Relevance (MMR) application



For our original dataset (17,018 articles) we conducted topic modeling to explore further the prevalent topics in the dataset's content.
We utlized BERTopic which identified a total of 81 distinct topics. After a manual review and merging topics with high similarity based on their most representative words we consolidated the list down to 60 topics. Moreover, 4,835 articles were not included in any specific topic and were thus excluded from further analysis.

In BERTopic each topic is assigned a unique number as identifier,indicating its rank based on the number of associated articles - the topic ‘0’ contains the most articles, whereas the topic ‘60’ contains the fewest. Additionally, each topic is characterized by the four most representative words. Therefore, to represent each topic name we keep both the identifier and the 4 most representative words.

Below are the top10 topics are derived from the model:

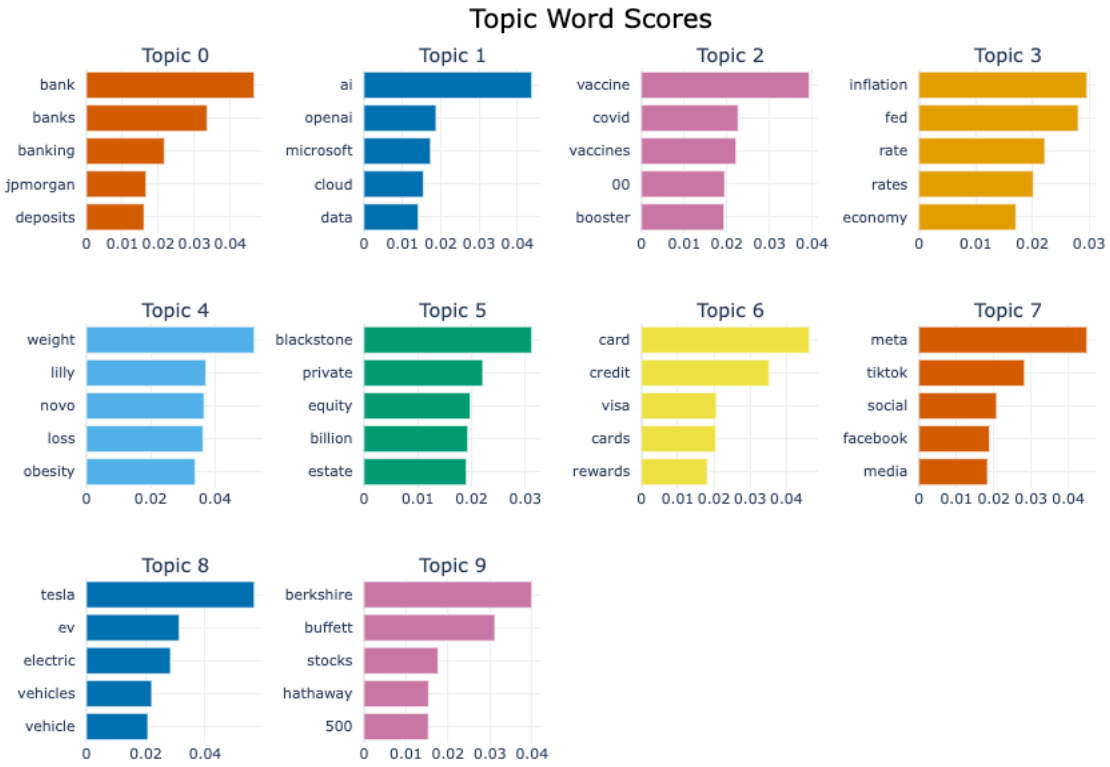


Fig 1.5.1: Representation of top 10 topics by the five most representative words

For the top 5 articles, related to banks, technology, vaccines, inflation, and health care, respectively, we see how these topics have evolved over the past 5 years.

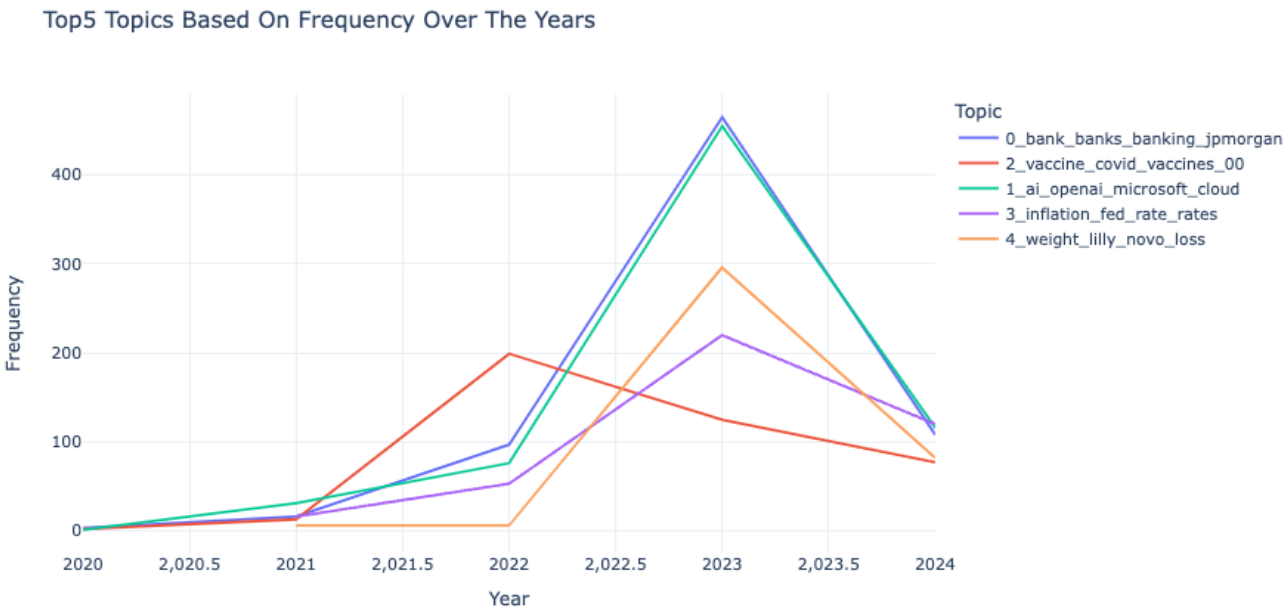


Fig1.5.2: Evolution of the 5 most discussed topics from 2020 to February 2024.

During the first two months of 2024, most articles focus on economic issues, with emphasis on inflation and Federal Reserve Rates, banks, and technology. This trend in financial and technological reporting is a continuation from the previous year when such topics comprised the majority of articles. Backtracking to 2022, it is notable that content related to covid-19 and vaccines was significantly more common, showing how the global health situation affected the news.

Furthermore, we identify the most discussed topics within each sector (Fig:1.5.3). In the financial sector, the focus was on banks and private equity. For health care, the most notable topics are the weight loss drug developed by Eli Lilly and Covid-19 vaccines. In the technology sector, artificial intelligence, particularly in relation to companies such as openAI and Microsoft emerged as the most prevalent subject, followed by developments in social media platforms linked to companies such as Meta.

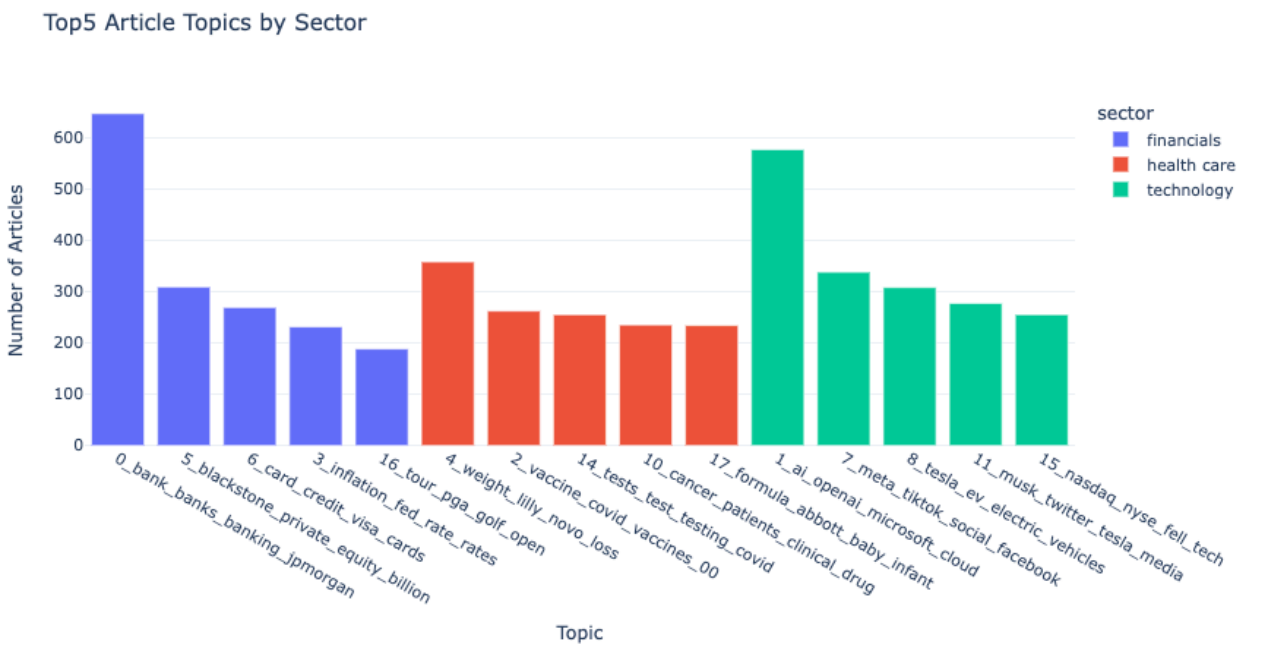


Fig1.5.3: The five most discussed topics within each sector: financials, health care, technology

Following that, we explore the top 10 topics based on their overall frequency in our dataset, focusing solely on the content as these topics are not represented by individual titles.

First, from Fig1.5.4 we observe that all topics have wide ranges, indicating high variability in the articles, with a range from very positive to very negative sentiments. However, the median sentiments are close to zero, with healthcare articles showing slightly higher median sentiments. Topics related to artificial intelligence and COVID-19 vaccinations show the most positive and negative outliers, respectively, suggesting a prevalence of articles with polarizing content.

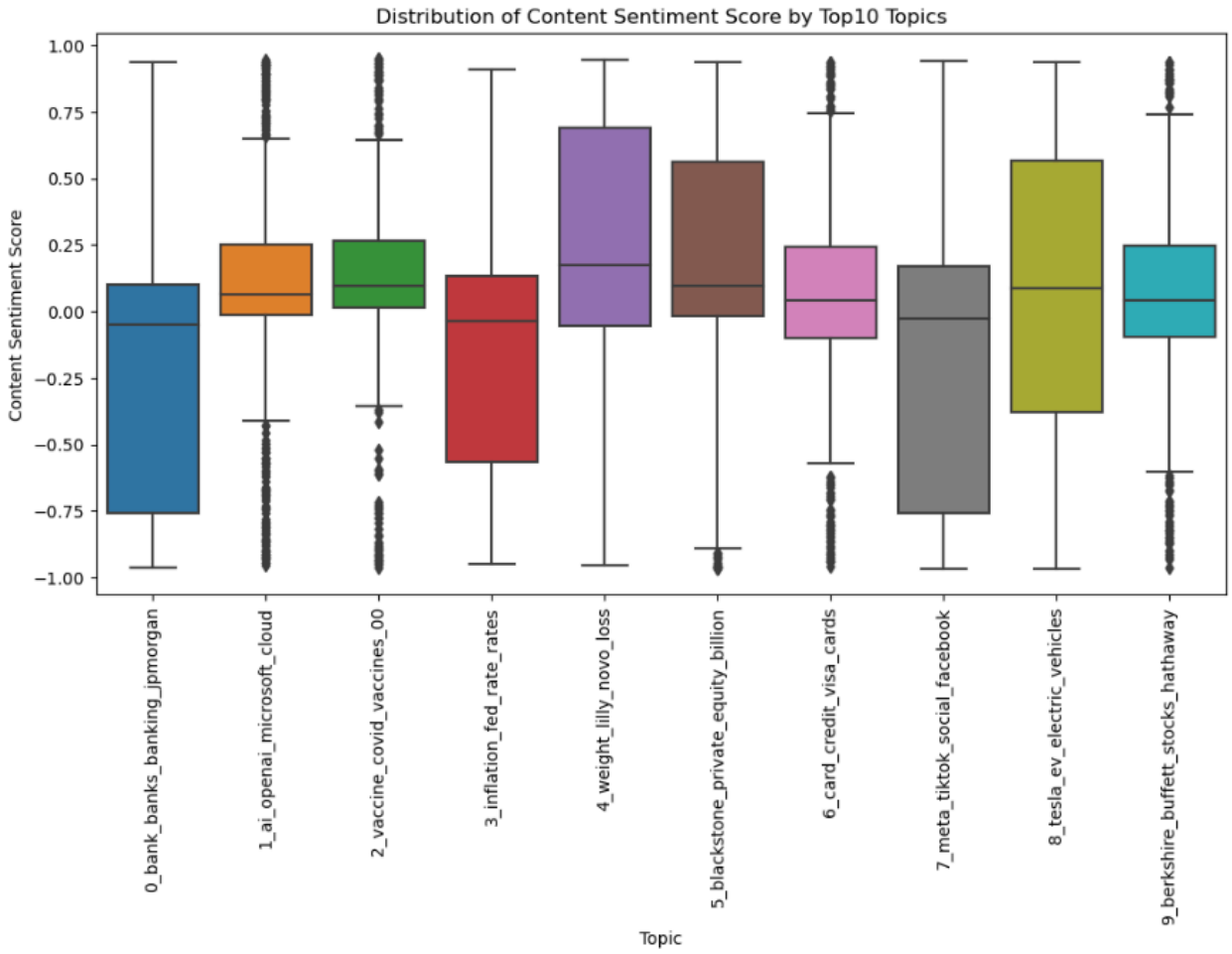


Fig1.5.4: Distribution of Content Sentiment Scores for the ten most frequent topics

Now we focus on the correlation between two-day returns and content sentiments for each topic. The figure below displays these results. Articles related to Tesla and electric vehicles show the highest positive correlation at 0.3, while those concerning credit cards show a slight negative correlation of -0.04. However, other financial topics such as banks, private equity, Warren Buffet's company, and stocks generally seem to have higher correlations compared to technological topics like AI.

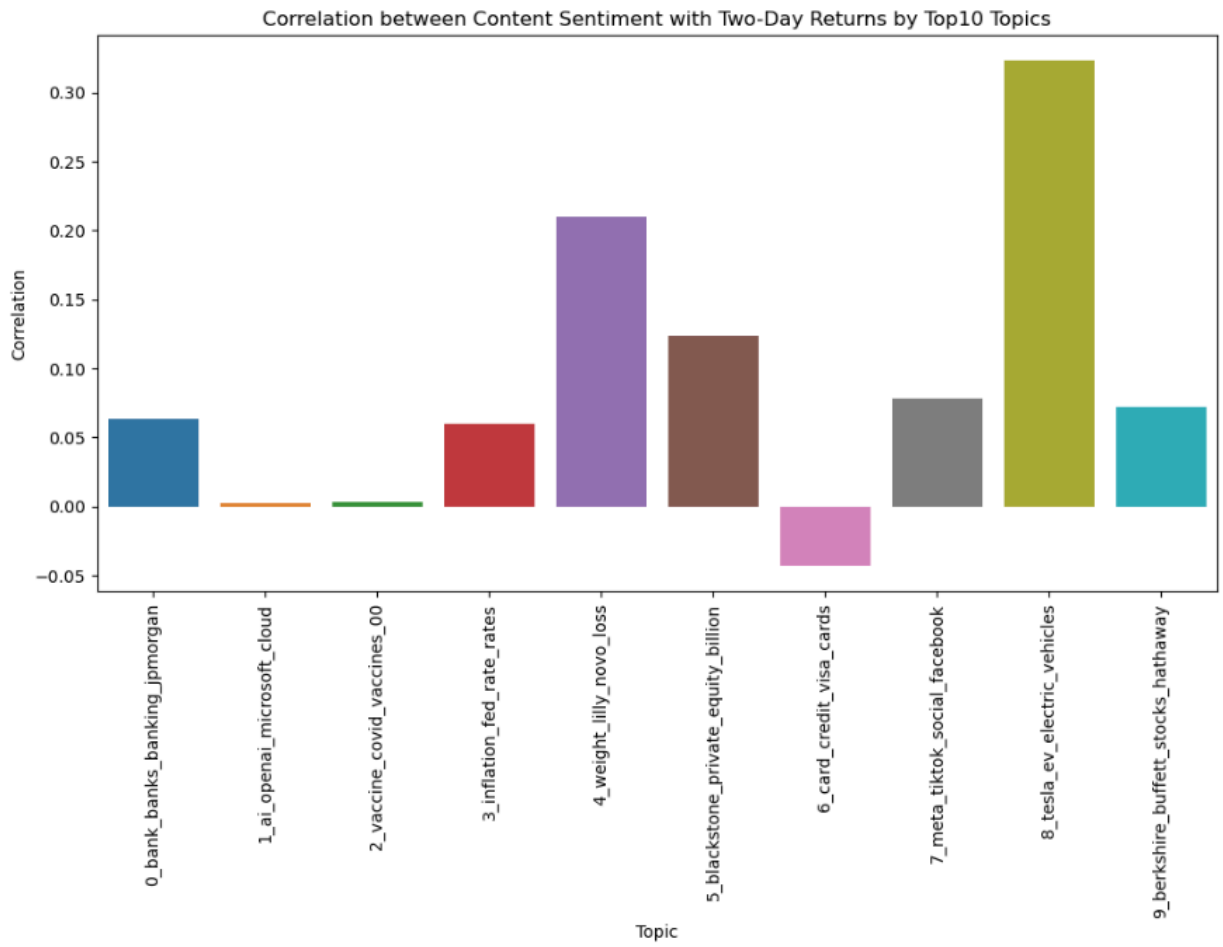


Fig1.5.5: Distribution of Content Sentiment Scores for the ten most frequent topics

To conclude, it is important to note that this analysis covers only the top 10 topics, and does not necessarily reflect the broader patterns across all sectors. Articles on technological topics may relate to companies in the financial or healthcare sector, and vice versa.

Section 2 Sentiment Factor Modeling

In this part, we are trying to use the sentiment score we got from the previous section to trade among all 30 stocks or each sector respectively, and conduct backtesting to see how they perform.

2.1 Build up trading strategy based on sentiment factor

In the previous section, we obtained the sentiment scores. For stock A on day T, there may be N related articles. We define the sentiment score of stock A on day T as the sum of sentiment scores for all N articles related to stock A on that day, denoted as

Sentiment score A,T = $\sum_{i=1}^{N}$ sentiment score A,T,i,

where sentiment score A,T,i represents the sentiment score of the i-th article about stock A on day T. We assume there is a lag in the impact of financial news on returns, denoted as t0. Additionally, we consider that news within a certain time range affects the returns over T days, with the length of this time range denoted as t1. Specifically, news within the time range [T-t0-t1, T-t0] affects the returns over T days. Within this time range, we assume that the impact of sentiment scores on returns remains consistent each day.

When the sentiment score of a particular stock exceeds a threshold value θ, we execute a trade for that stock. Specifically, when $sentiment\ score > \theta 1$, we interpret the market sentiment as bullish for the stock, anticipating an increase in the stock price, and thus we buy the stock. Conversely, when $sentiment\ score <- \theta 2$, , we interpret the market sentiment as bearish for the stock, expecting a decrease in the stock price, and therefore we sell the stock.

For simplicity, we trade each stock separately. For each stock, the initial capital is $10,000, so the total initial capital is $300,000. When executing a buy operation for a certain stock, the transaction amount is the minimum of 10% of the initial capital for that stock and the remaining cash flow for that stock. When executing a sell operation for a certain stock, the transaction amount is the minimum of 10% of the initial capital for that stock and the market value of the remaining holding for that stock. This approach helps to avoid overly aggressive trading.

Therefore, our trades involve three parameters: shift days t0, rolling days t1 and trading threshold $\theta 1$ and $\theta 2$. In the next part, we will optimize these three parameters.
After finding the best parameters, we will use the strategies using these parameters to conduct bactest and see the performances.

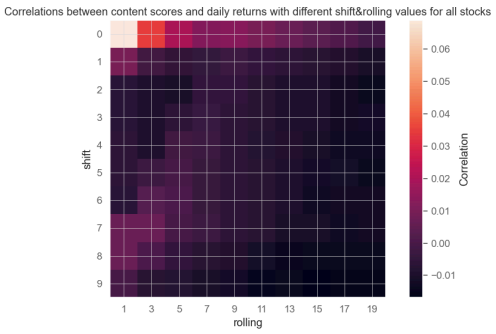
2.2 Optimization among 4 hyperparameters

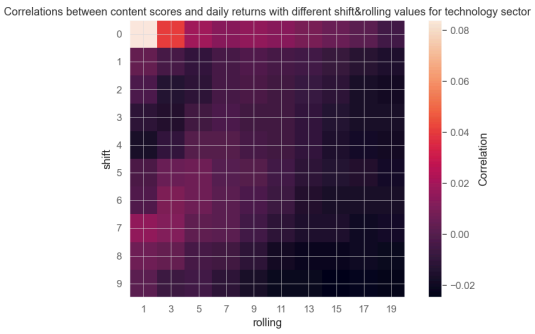
To adjust the parameters, we divide the entire dataset (from 2020-08-10 to 2024-02-23, totaling 890 trading days) into two parts: from 2020-08-10 to 2022-11-01 (a total of 562 trading days) and from 2022-11-01 to 2024-02-23 (a total of 328 trading days). The former part is used to search for the optimal parameter combinations, while the latter part is used for backtesting.

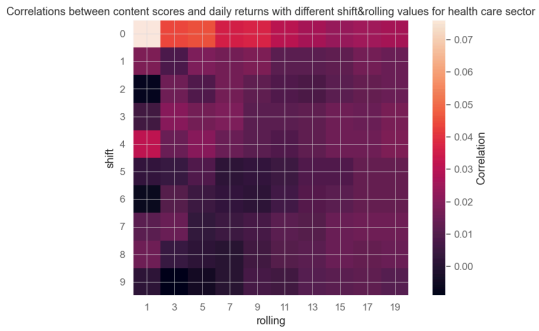
We search for the best parameter combinations among shift days $t0 \in N \cap [0, 9]$, rolling windows $t1 \in Nodd \cap [1, 20]$, $\theta 1 \in 0.05 * [0, 9] \cap N$, and $\theta 2 \in -0.05 * [0, 9] \cap N$.

2.2.1 Shift days and rolling window

To determine the optimal values for shift days and rolling window, we create a heatmap plot showing the correlation between sentiment scores and price returns for different combinations of these values.







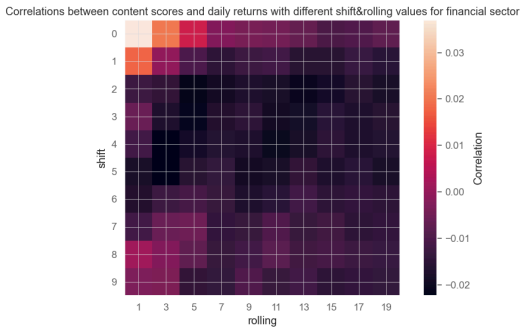
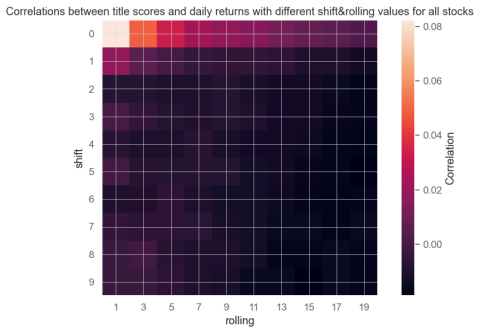
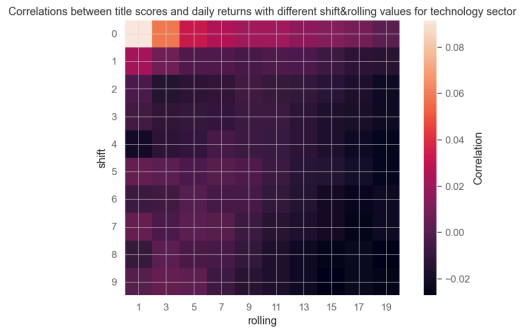
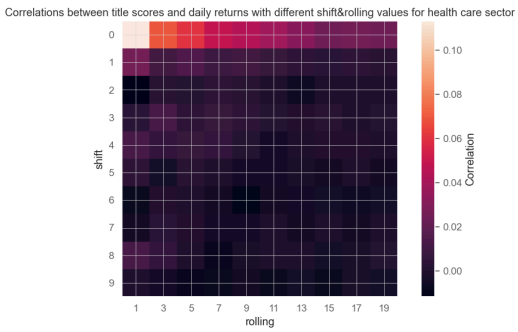


Fig 2.2.1: Correlation between content scores and daily returns for all stocks, technology sector, health care sector and financial sector







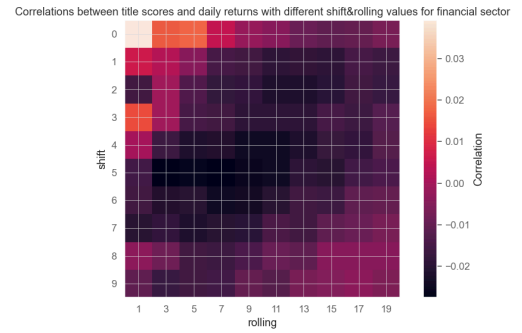


Fig 2.2.2: Correlation between title scores and daily returns for all stocks, technology sector, health care sector and financial sector

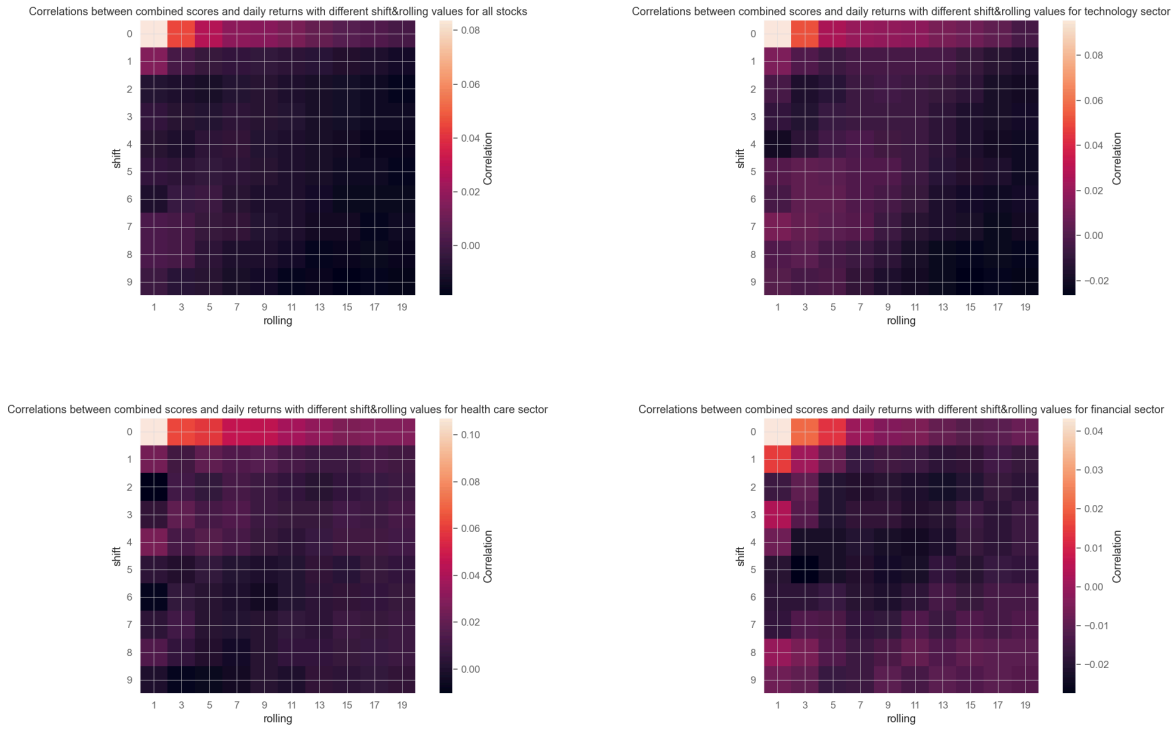


Fig 2.2.3: Correlation between combined scores and daily returns for all stocks, technology sector, health care sector and financial sector

We choose shift days $t0 = 0$, and rolling windows $t1 = 1$ for all 30 stocks and each sector for the content scores, title scores and combined scores. Namely, we use the original sentiment score without shift and rolling operations.

2.2.3 Thresholds θ1 and θ2

To determine the optimal values for thresholds $\theta 1$ and $\theta 2$, we create a heatmap plot showing the sharpe ratios by strategies with different $\theta 1$ and $\theta 2$ values.

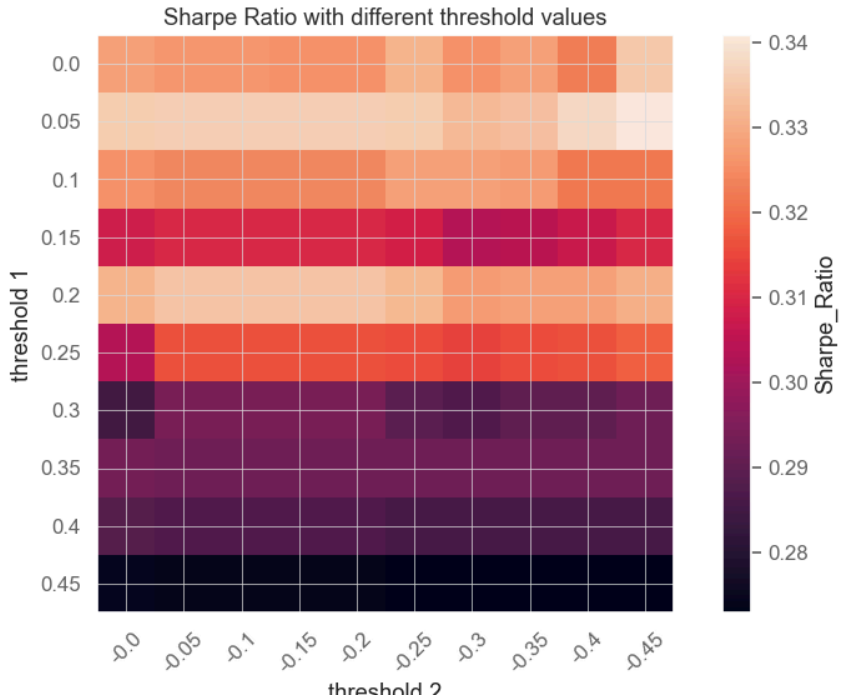


Fig 2.2.4: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ of content strategy for all 30 stocks

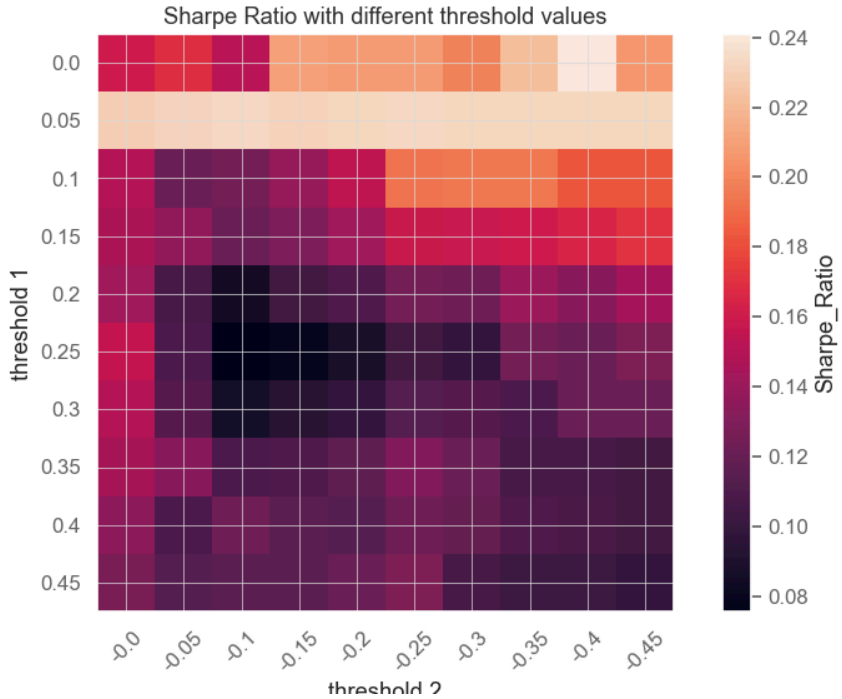


Fig 2.2.5: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ of title strategy for all 30 stocks

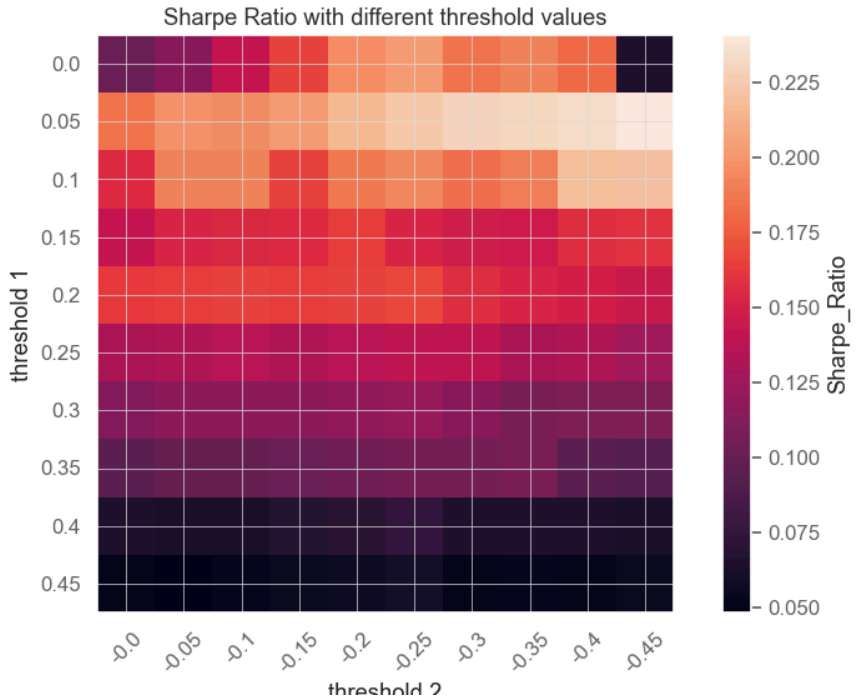


Fig 2.2.6: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ of combined strategy for all 30 stocks

We can even use more slight intervals to find better combinations of thresholds based on the areas with relatively high sharpe ratio we derived from Fig. 2.2.4-Fig.2.2.6.

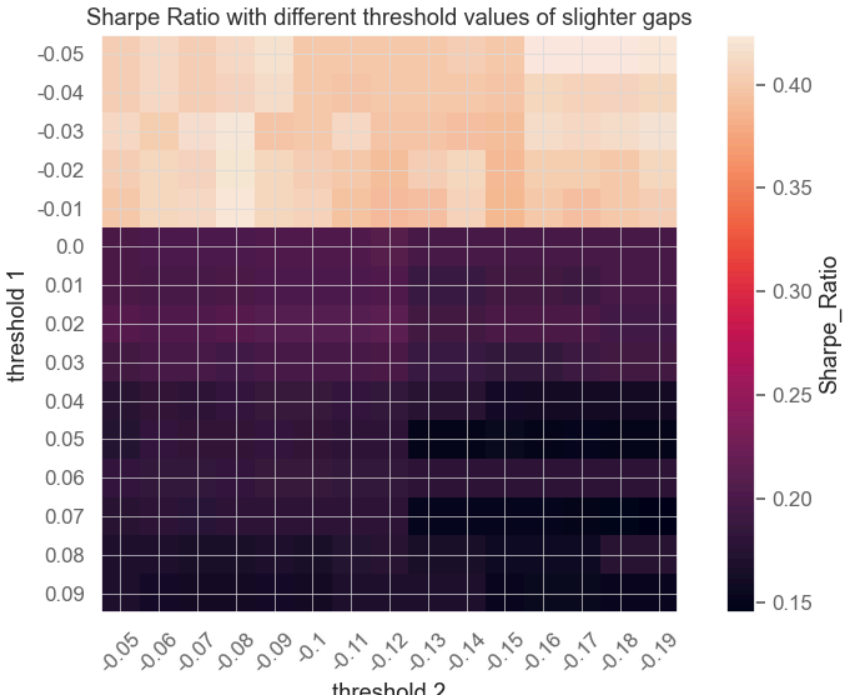


Fig 2.2.7: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ given slighter gaps of content strategy

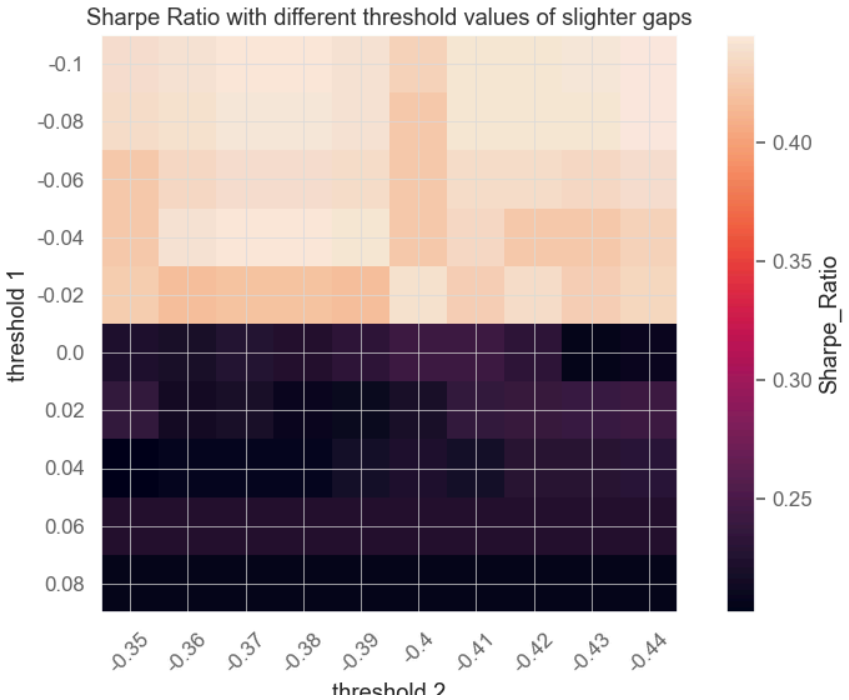


Fig 2.2.8: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ given slighter gaps of title strategy

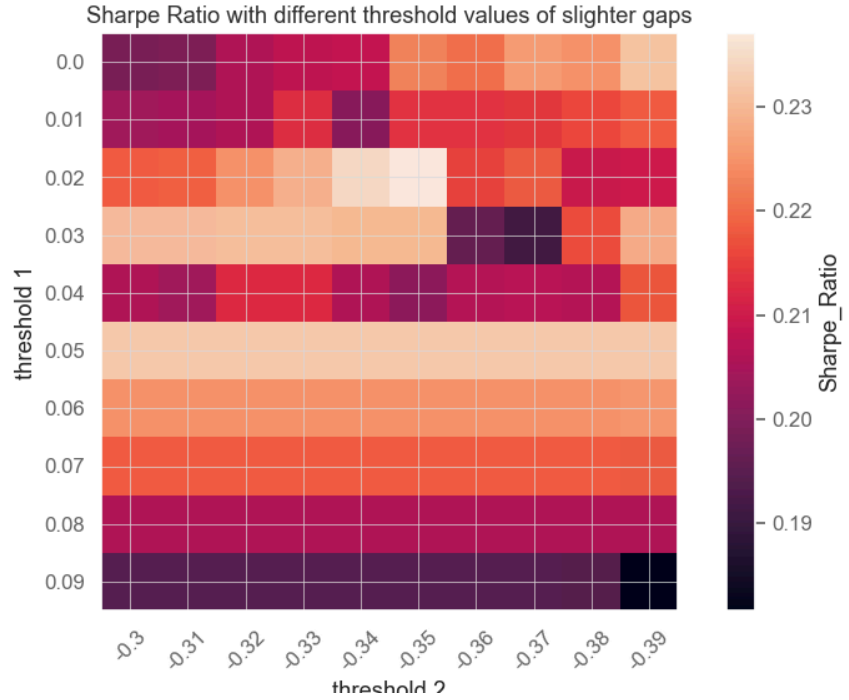


Fig 2.2.9: Sharpe Ratio with different thresholds $\theta 1$ and $\theta 2$ given slighter gaps of combined strategy

From Fig. 2.2.7, When $\theta 1$ decreases from 0 to -0.1, we observe a significant leap in the Sharpe ratio. When $\theta 1$ stays below -0.1, the Sharpe ratio remains relatively stable with variations in $\theta 1$ and $\theta 2$. Also, a slowly increasing trend in Sharpe Ratio could be seen as the $\theta 2$ increases in this area. We select a more suitable parameter combination within this range for the trading strategy based on content sentiment, which is $\theta 1 = -0.04, \theta 2 = -0.08$. Similarly, we choose $\theta 1 = -0.08, \theta 2 = -0.37$ for title sentiment from Fig. 2.2.8.

From Fig.2.2.9, the Sharpe Ratio stays stable as $\theta 2$ changes when $\theta 1 > 0.05$. Within this range, we choose $\theta 1 = -0.06, \theta 2 = -0.34$.

Sentiment	$\theta 1$	$\theta 2$
Content	-0.04	-0.08
Title	-0.08	-0.37
Combined(Title-Content)	-0.06	-0.34

Table. 2.2.1: Parameters chosen for the three kinds of sentiment among 30 stocks

At first glance, it may seem unreasonable that the threshold for buying stocks is negative, but from Fig 1.2.17 - Fig 1.2.19, we can glimpse some insights. 'Somewhat Negative' sentiment category can also have positive correlations to two-day return, which suggests that the stocks with slightly negative sentiment scores can even have an upcoming increase in return.

2.2.3 Out of sample test using parameters we found

Applying the parameters combinations we found in the previous subsection, we conduct out of sample test on the test set (from 2022-11-01 to 2024-02-22).

The results are shown in the table below:

Table 2.3.1 The results of out of sample test

2This result may seem promising at first glance, but we can notice that the results obtained from the out-of-sample test are better than those from the training data, indicating some potential issues. Firstly, limited text data may lead to insufficient training and out-of-sample test data, resulting in less robust results overall. Secondly, the data used is a time series, and price returns may be correlated with time. The training data is closer to the outbreak of the COVID-19 pandemic, while the out-of-sample test data is further away, potentially leading to different patterns in returns. Lastly, there are many areas for improvement in the trading strategy. For example, combining content score and title score with weights to

optimize factors, or allocating daily trading volume for each stock based on sentiment score weights. These measures could potentially improve out-of-sample performance.

2.3 Numerical data factor

Backtest:

First we construct the mean-reversion factor: add up the returns over the past i days, then take the inverse to get our original signal, then standardize the signal, conduct a simple backtest, calculate the Sharpe ratio, and draw the Sharpe ratio and the image of parameter i（rolling value）, roughly determine its optimal parameter range. The image performs well and is stable near i=7, 30, and 70.

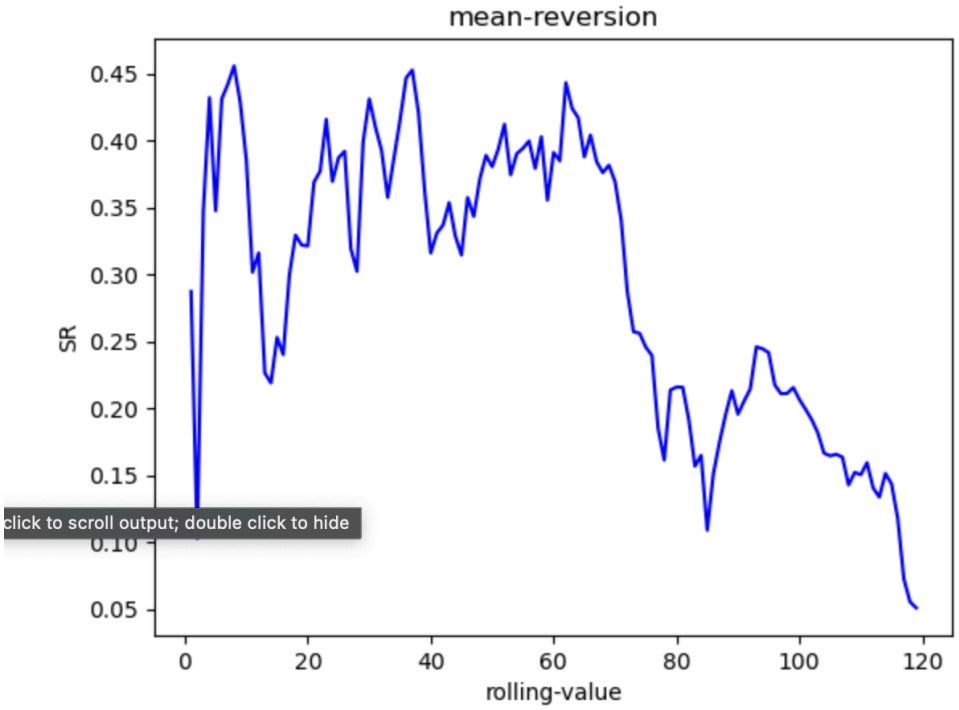


Fig 2.3.1: Sharpe Ratio with different rolling-value of mean-reversion factor

Next, in the same way, we construct the momentum factor: first shift to 253-i days ago, then sum the returns of i days before 253-i days to get our original signal, and then standardize the signal, and Conduct a simple backtest, calculate the Sharpe ratio, and draw the image of the Sharpe ratio and parameter i (rolling-value) to roughly determine the optimal parameter range. The image performs well and is stable around i=16,180.

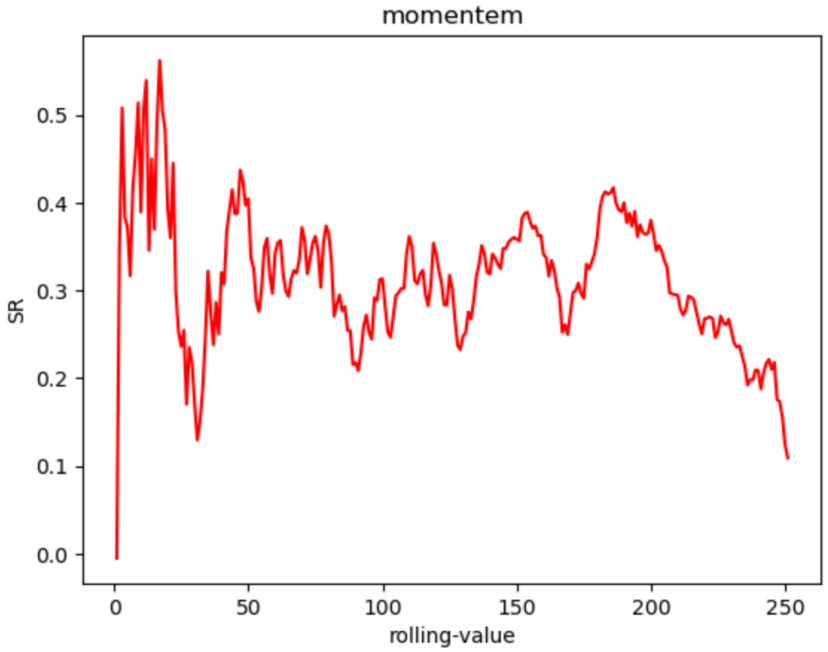


Fig 2.3.2: Sharpe Ratio with different rolling-value of momentum factor

Next, in the same way, we construct the trading volume factor: add up the trading volume in the past i days, and then get our original signal, then standardize the signal, conduct simple backtesting, calculate the Sharpe ratio, and draw Get the image of Sharpe ratio and parameter i (rolling-value), and roughly determine the optimal parameter range. The image performs well and is stable around i=25,150.

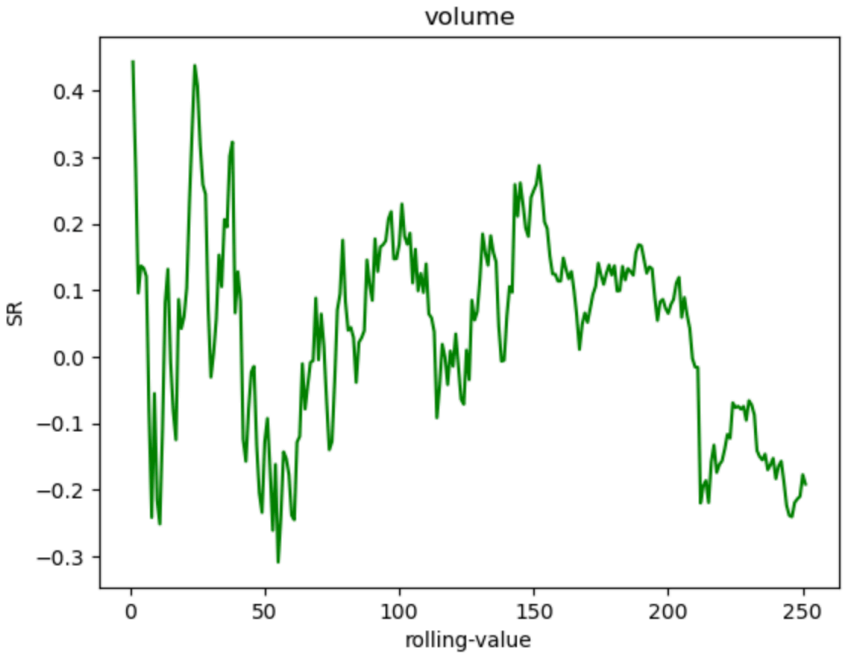


Fig 2.3.3: Sharpe Ratio with different rolling-value of volume factor

After preliminary parameter analysis, we formulate further strategies, we call it backtest 1: set thresholds and other conditions, and manage the generated signals. And draw the Sharpe ratio image obtained by different factors as the threshold value changes (a-value) as follows.

The first is the image of mean-reversion. We can see that as the threshold setting(a-value) increases, the Sharpe ratio shows a downward trend, which means that this factor generates a sensitive signal and is more suitable for higher-frequency trading strategies.

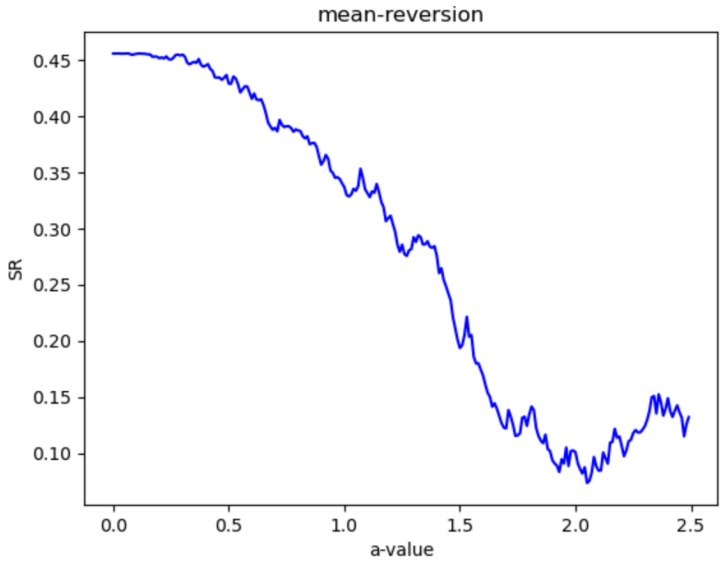


Fig 2.3.4: Sharpe Ratio with different a-value(threshold) of volume factor

Next is the image of momentum. We can see that as the threshold setting increases, its Sharpe ratio shows an upward trend, which means that the factor generates a conservative signal and is more suitable for lower-frequency trading strategies.

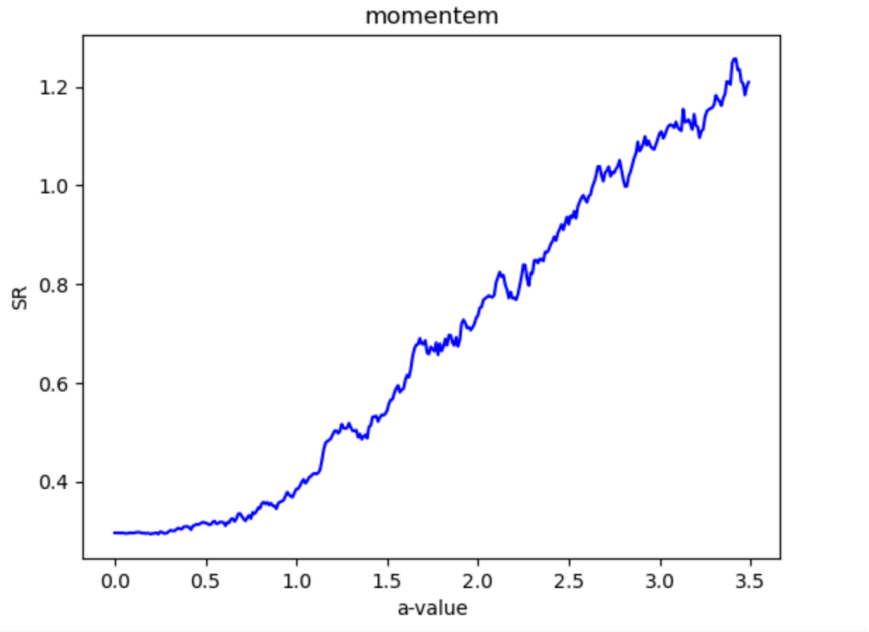


Fig 2.3.5: Sharpe Ratio with different a-value(threshold) of momentum factor

Then there is the image of volume. We can see that as the threshold setting increases, its Sharpe ratio first increases and then decreases. This means that the factor generates a moderate signal.

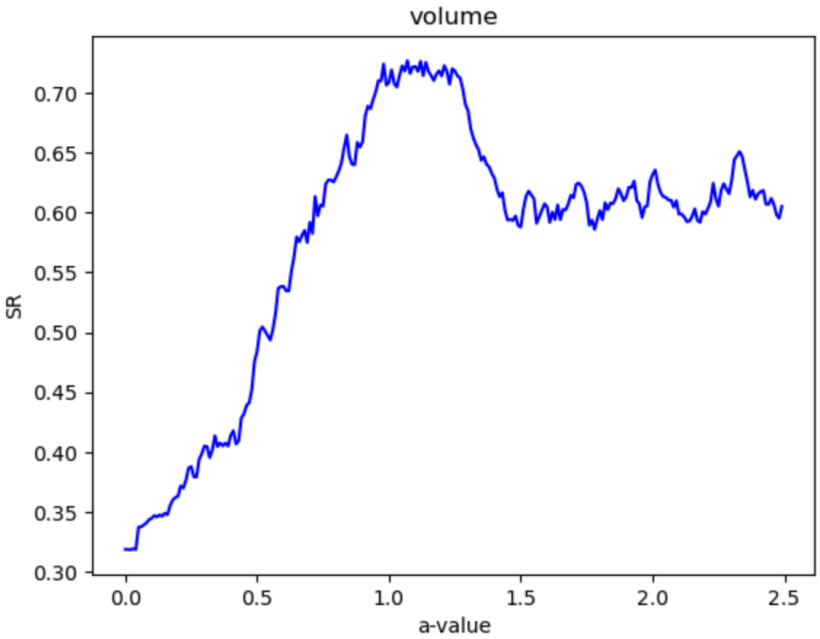


Fig 2.3.6: Sharpe Ratio with different a-value(threshold) of volume factor

In addition, we also performed group backtesting for each factor, we call it backtest 2 : each time, we took the n stocks with the largest factor signal value and performed long, and the n stocks with the smallest factor signal value performed short, looked at the Sharpe ratio of the resulting transactions, and drew The images of Sharpe ratio and n-value are as follows:

The first is mean-reversion. We find that when using this factor to invest, as the n-value of the number of stocks in the portfolio increases, the Sharpe ratio of the factor shows an upward trend, which means that when using this factor to invest, you should choose as many options as possible. Only stocks are used to hedge risks, because this is a relatively sensitive factor.

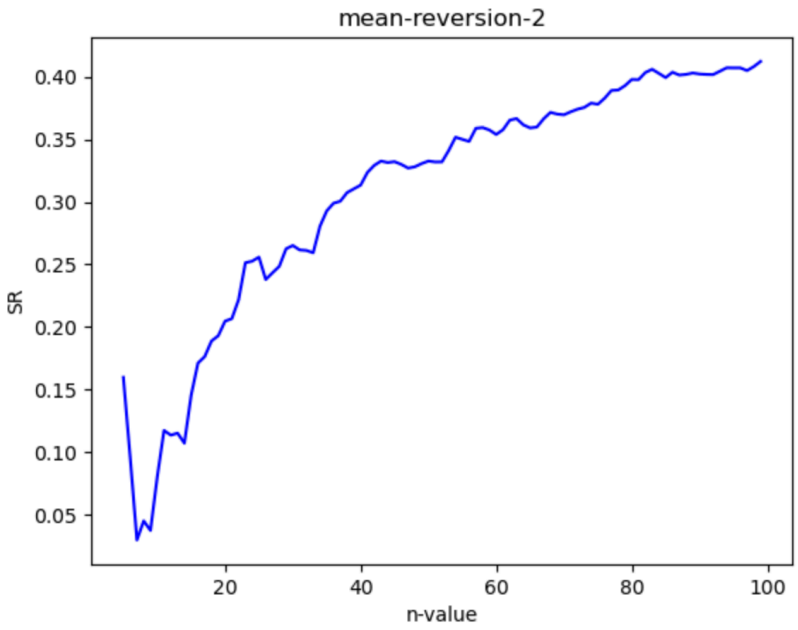


Fig 2.3.7: Sharpe Ratio with different n-value(number of stocks in portfolio) of mean-reversion factor

The second is the momentum factor. We found that when using this factor to invest, as the n-value of the number of stocks in the portfolio increases, the Sharpe ratio of the factor shows a downward trend, which means that when investing using this factor, you should try your best to Select fewer stocks because this is a relatively conservative factor. The stocks with high signal values derived from it have an absolute advantage, and there is no need to add extra stocks to the investment portfolio.

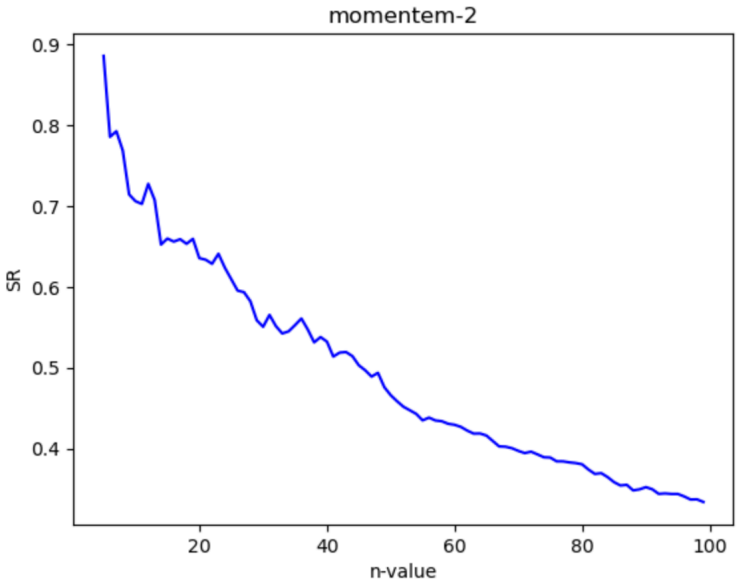


Fig 2.3.8: Sharpe Ratio with different n-value(number of stocks in portfolio) of momentum factor

Then there is the image of volume. We find that when using this factor to invest, as the n-value of the number of stocks in the portfolio increases, the Sharpe ratio of the factor first increases and then decreases, which means that when investing using this factor, it should Select stocks appropriately as this is a relatively moderate factor.

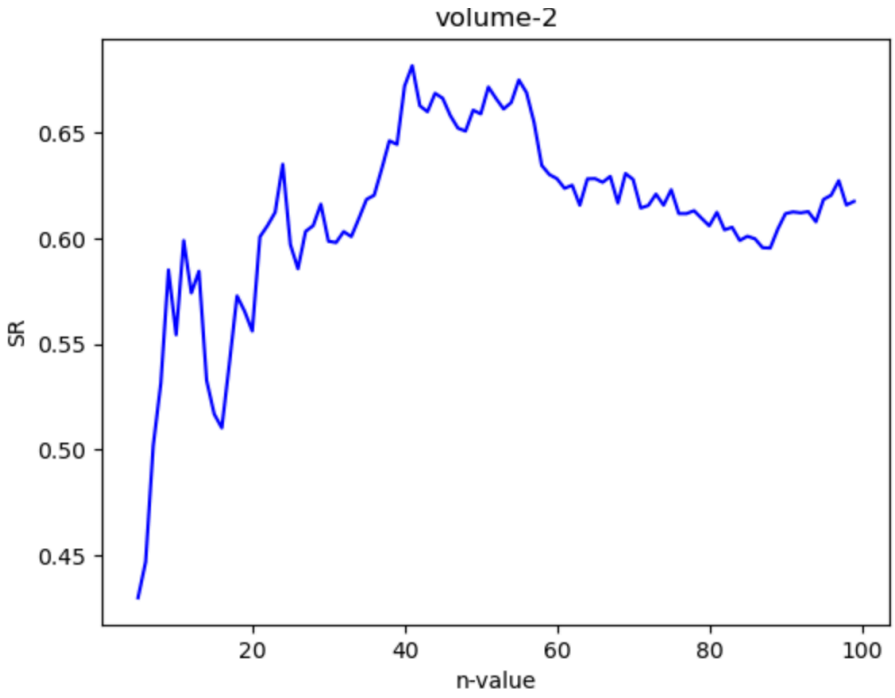


Fig 2.3.9: Sharpe Ratio with different n-value(number of stocks in portfolio) of volume factor

The above results of the second group backtest and the first backtest are consistent in the essential analysis of factors, and also verify our ideas.

2.4 Test and Analyze Results

After the backtest is completed, we need to test the performance of the factors on the test set. First, we use linear regression to test each factor:

First perform linear regression on the mean-reversion factor

This is our fitting result. Judging from the picture, the effect is relatively good. The obtained MSE is only 0.0001, R^2 score: 0.1884

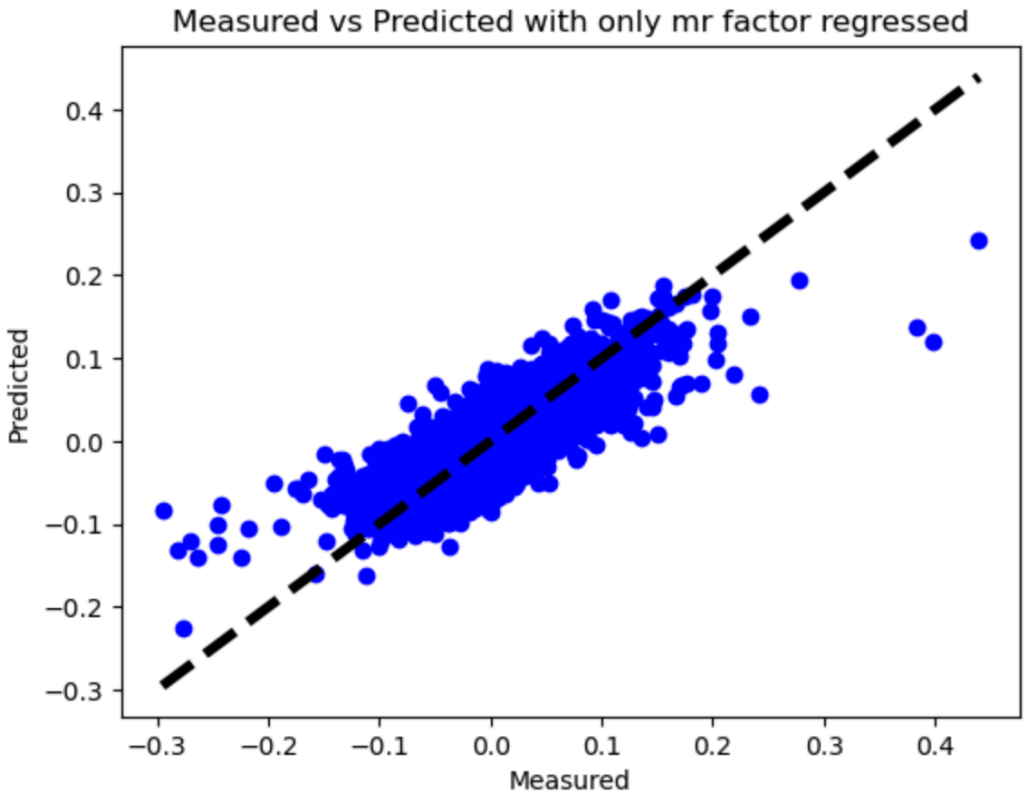


Fig 2.4.1: linear regression of mean-reversion factor

Next is the linear regression of the momentum factor. From the figure, the regression effect is also good, mse is only 0.0001, R^2 score: 0.2521

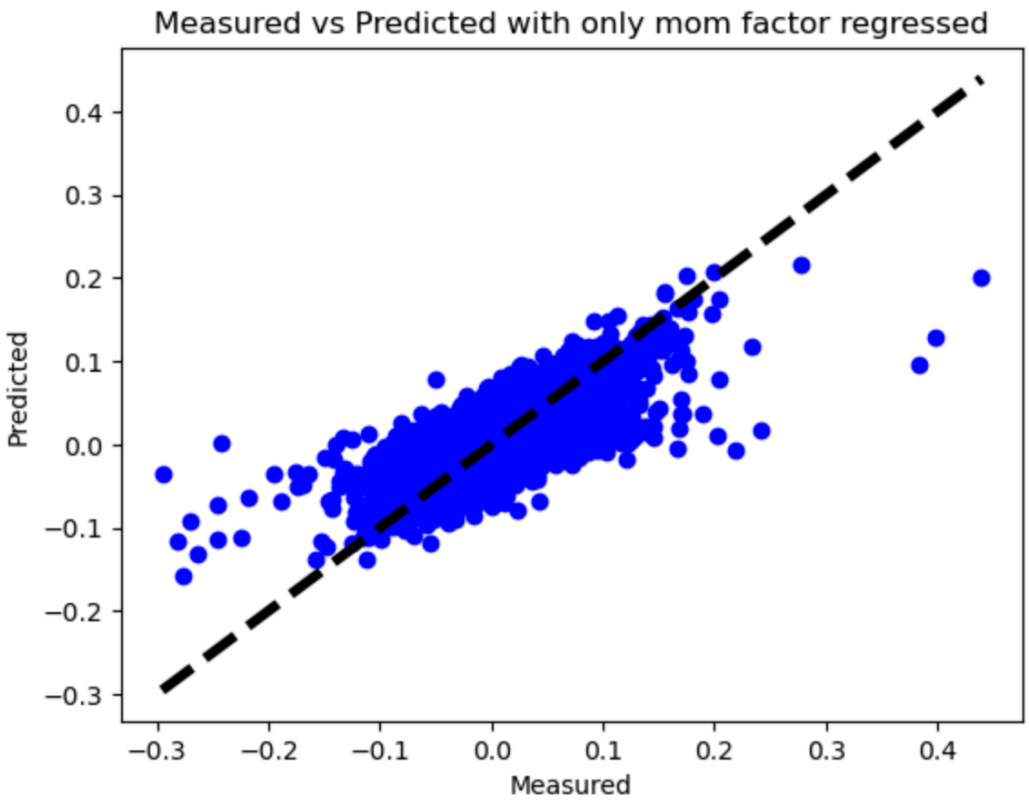


Fig 2.4.2: linear regression of momentum factor

Also, there is the linear regression of the volume factor. From the picture, the effect is not ideal, so for the time being, we will consider other factors first.

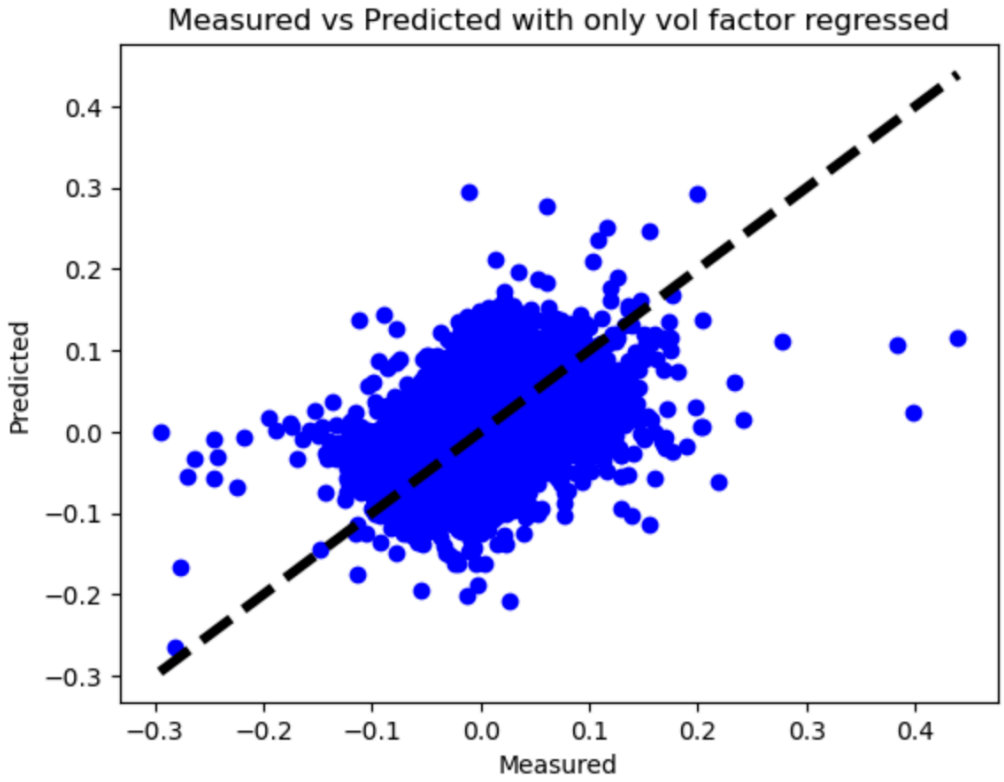


Fig 2.4.3: linear regression of volume factor

Finally, there is the linear regression of sentiment scores factor. From the picture, the effect is very good. The R^2 score reaches: 0.7340

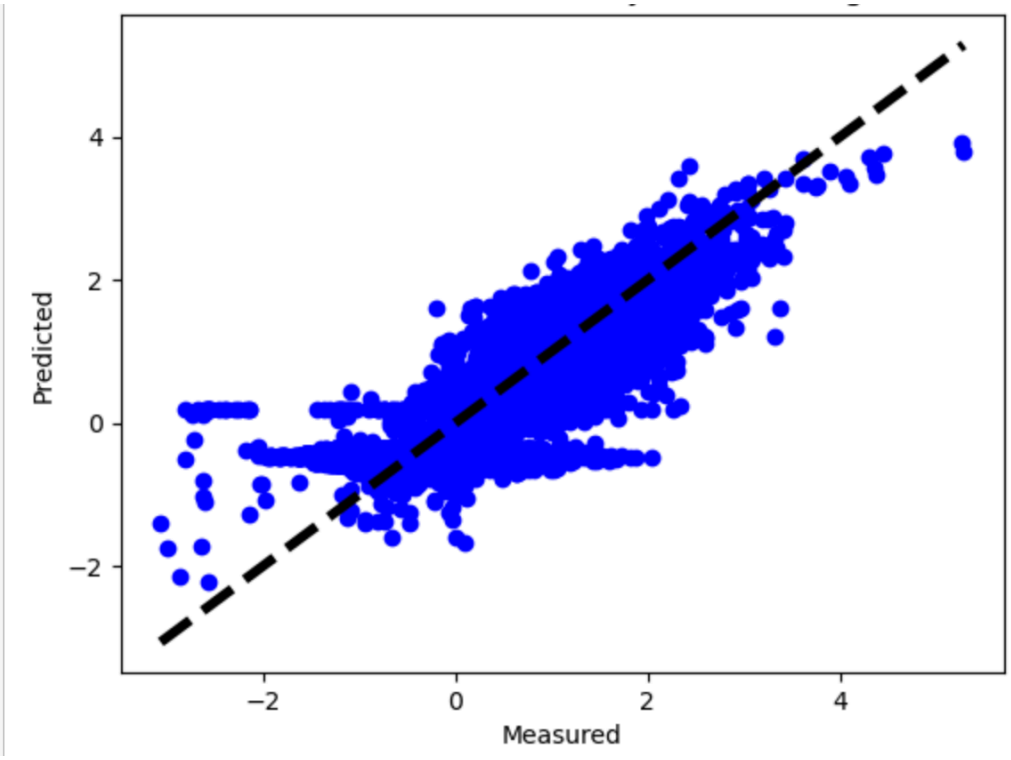


Fig 2.4.4: linear regression of sentiment scores factor

After the single-factor regression is completed, it is combined into a multi-factor form for combined weighted regression and the results are as follows:

The linear regression result with sentiment scores is:

R^2 score: 0.7652

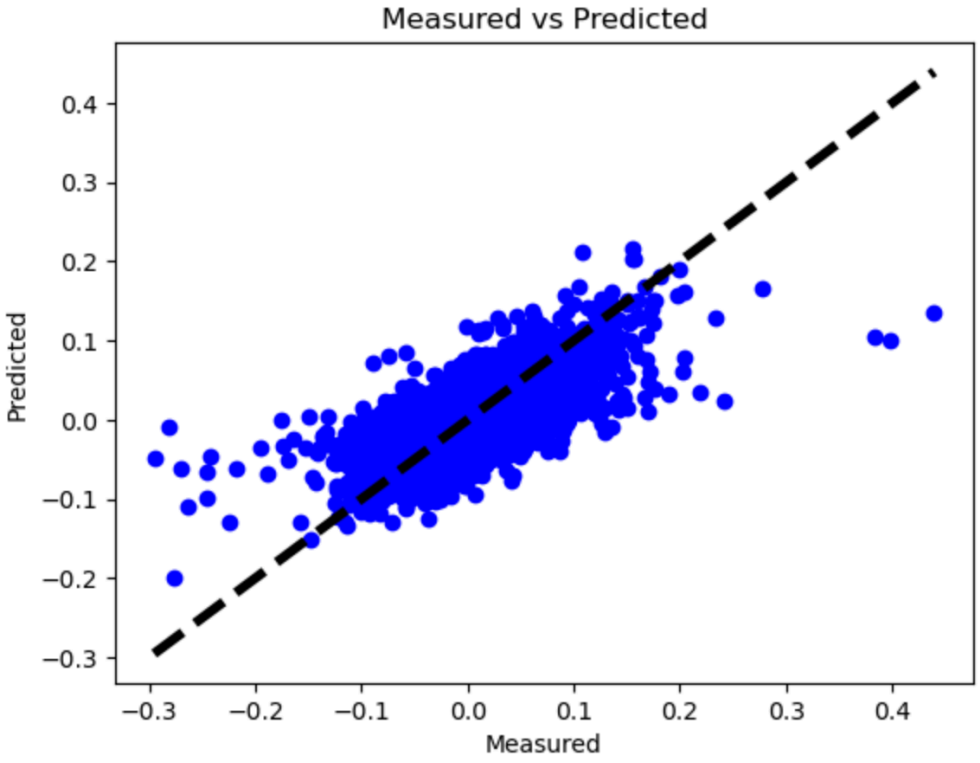


Fig 2.4.5: linear regression result with sentiment scores

From this we can judge that the sentiment score factor has relatively strong interpretability and accuracy.

Conclusions and Future Work

In our project we examined the relationship between articles sentiments and two-day return windows, leveraging FinBERT for sentiment analysis. Our results indicated sufficient correlation between content and title sentiments across the three sectors - technology, financial, healthcare - where similar trends and some distinct characteristics were prevalent. These findings led us to incorporate sentiments as an additional factor for trading pursposes.

However, there are still some deficiencies in this study. Although we utilized a state-of-the-art Large Language Model (LLM) specialized in financial corpora, further fine tuning is required for more reliable sentiment detection. Therefore, in ongoing efforts to refine sentiment analysis within the financial context, we plan to manually label a substantial dataset to fine-tune FinBERT to predict sentiment scores with increased accuracy. Furthermore, although we tried to mitigate potential inaccuracies from the FinBERT model, by focusing on indirect methods such as weighting more heavily or assigning scores only to the most frequently mentioned companies, we recognize the need for more direct improvements. In future work we intend to implement aspect-based sentiment analysis to accurately capture the sentiments expressed about individual companies within articles.

Additionally, to go one step further another target is to explore the correlation between sentiments and market volatility, as understanding this relationship could significantly improve trading strategies and risk management techniques.

Bibliography

[1] https://gnews.io/

[2] https://www.diva-portal.org/smash/get/diva2:1636643/FULLTEXT01.pdf

[3] https://medium.com/@n83072/topic-modeling-bertopic-ca1b73a035f2

[4] https://arxiv.org/abs/1908.10063

[5] https://medium.com/@nick-tan/topic-modeling-with-bertopic-a-cookbook-with-an-end-to-end-example-part-1-3ef739b8d9f8

[6] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3910214

[7] https://www.mdpi.com/2079-9292/12/12/2605

[8]https://medium.com/data-reply-it-datatech/bertopic-topic-modeling-as-you-have-never-seen-it-before-abb48bbab2b2

[9] https://wandb.ai/ivangoncharov/FinBERT_Sentiment_Analysis_Project/reports/Financial-Sentiment-Analysis-on-Stock-Market-Headlines-With-FinBERT-HuggingFace--VmlldzoxMDQ4NjM0

[10] https://medium.com/@priyatoshanand/handle-long-text-corpus-for-bert-model-3c85248214aa

[11]https://www.ijcai.org/proceedings/2020/0622.pdf