Can ChatGPT Compute Trustworthy Sentiment Scores from Bloomberg Market Wraps?

Baptiste Lefort, Eric Benhamou, Jean-Jacques Ohana, David Saltiel, Beatrice Guez, D Challet

Can ChatGPT Compute Trustworthy Sentiment Scores from Bloomberg Market Wraps?

B. Lefort[1,2], E. Benhamou[1,3], JJ. Ohana[1], D. Saltiel[1], B. Guez[1], D. Challet[2]
[1]: Ai for Alpha, [2]: CentraleSupelec, [3]: Dauphine PSL
{baptiste.lefort, eric.benhamou, jean-jacques.ohana, david.saltiel, beatrice.guez}@aiforalpha.com
damien.challet@centralesupelec.fr

Abstract

We used a dataset of daily Bloomberg Financial Market Summaries from 2010 to 2023, reposted on large financial media, to determine how global news headlines may affect stock market movements using ChatGPT and a two-stage prompt approach. We document a statistically significant positive correlation between the sentiment score and future equity market returns over short to medium term, which reverts to a negative correlation over longer horizons. Validation of this correlation pattern across multiple equity markets indicates its robustness across equity regions and resilience to non-linearity, evidenced by comparison of Pearson and Spearman correlations. Finally, we provide an estimate of the optimal horizon that strikes a balance between reactivity to new information and correlation.

Keywords: sentiment analysis, ChatGPT, stock exchange, financial news

1. Introduction

Finance has a longstanding tradition of employing Natural Language Processing (NLP) to extract valuable insights from textual data and news (Tetlock, 2007; Schumaker and Chen, 2009). The financial world has always been at the forefront of embracing technological innovation. From the inception of electronic trading to the burgeoning realm of fintech, financial services have undergone significant evolution, especially with the arrival of AI and ML technologies (Arner et al., 2015; Fatouros et al., 2023).

Sentiment analysis stands out as a cornerstone in this transformation (Poria et al., 2016). It plays a crucial role in deciphering market sentiments, offering invaluable predictive insights. Historically, the financial sector leaned on handpicked word lists and basic ML techniques for sentiment analysis (Tetlock, 2007; Schumaker and Chen, 2009). Yet, with NLP's rapid advancements, a slew of advanced methods has come to the fore. Models like BERT and its finance-centric sibling, FinBERT, have elevated sentiment analysis's precision (Devlin et al., 2018; Liu et al., 2021).

However, the financial realm brings its set of challenges for sentiment analysis (Loughran and McDonald, 2011). Financial news is a complex mesh of domain-specific jargon and layered emotions. A singular piece of news might carry different sentiments for multiple financial entities, making general sentiment analysis tools potentially misleading. News may also come after the facts and hence have no real predictive power. Furthermore, these tools often struggle with context-specific outputs, making them less versatile in diverse scenarios (Poria et al., 2017). Indeed, undertaking natural language processing (NLP) in finance is notably challenging due to the specificity of the corpus, as evidenced by diverse studies on financial texts, sentiment lexicons, and financial reports across various languages and financial systems (Li et al., 2022; Moreno-Ortiz et al., 2020; Ghaddar and Langlais, 2020) and can require knowledge graph (Oksanen et al., 2022) or language-specific corpus (Masson and Paroubek, 2020; Jabbari et al., 2020; Zmandar et al., 2022). Converting a sentiment score into an investment strategy is notably difficult (Yuan et al., 2020; Iordache et al., 2022)

With the advent of Large Language Models (LLMs), an AI paradigm has emerged with transformative potential (George and George, 2023). GPT, particularly its conversational variant, ChatGPT, has shown promise in refining financial applications (OpenAI, 2023). By leveraging ChatGPT's prowess in language comprehension, financial entities can enhance their sentiment analysis depth. This proficiency translates to better-informed investment decisions, optimized risk management, and more effective portfolio strategies. Furthermore, ChatGPT's capability to convey intricate financial insights in understandable terms makes it a potential game-changer in democratizing financial knowledge (Yue et al., 2023).

In this study, we design a sentiment analysis of Bloomberg markets wrap news using ChatGPT. Besides, we developed a two-step prompt-based process to extract information from text and convert this into a sentiment score. Finally, we show that this score enables us to understand better the effect of the news on the market especially regarding cyclic and counter-cyclic behavior. To sum up, the contributions of this paper are three folds:

1. We designed a two-step ChatGPT based sentiment analysis extraction from Bloomberg markets wrap news.

2. We proposed an index for assessing the ability of ChatGPT to give a sentiment to the news.

3. We demonstrated that this score reveals significant insights into market behavior and possesses robust predictive capabilities.

The rest of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 describes our prompt design and explains how using a two-step method for creating prompts can lead to better sentiment scores than using a one-step approach. Section 4 outlines the methodology for calculating the sentiment score. Section 5 evaluates the sentiment score validity. Section 6 discuss the trade-off between using short term predictions with lower correlation or longer period prediction but with the disadvantage of slow reaction to new informations. Section 7 review its robustness across various markets. Finally Section 8 concludes.

2. Related works

In the realm of finance and economics, several recent scholarly works have employed ChatGPT, such as Hansen and Kazinnik (2023), Cowen and Tabarrok (2023), Korinek (2023); Lopez-Lira and Tang (2023), and Noy and Zhang (2023). Hansen and Kazinnik (2023) elucidates how Large Language Models (LLMs) like ChatGPT can decipher Fedspeak, the nuanced language employed by the Federal Reserve to convey monetary policy decisions. Lopez-Lira and Tang (2023) explains proper prompting for forecasting stock returns. Both Cowen and Tabarrok (2023) and Korinek (2023) elaborate on ChatGPT's utility in economics education and research. Meanwhile, Noy and Zhang (2023) underscores ChatGPT's capability to augment productivity in professional writing tasks. Furthermore, Yang and Menczer (2023) showcases ChatGPT's aptitude for distinguishing credible news outlets.

Simultaneously, research by Xie et al. (2023) posits that ChatGPT's performance is comparable to rudimentary methods like linear regression for numerical data-based prediction tasks. Additionally, Ko and Lee (2023) endeavoured to employ ChatGPT in portfolio selection, albeit without discernible success. Our hypothesis attributes these varied outcomes to their reliance on historical numerical data for prediction, whereas ChatGPT's forte lies in textual tasks.

Our paper offers a novel perspective on this body of literature. It pioneers the assessment of ChatGPT's proficiency in forecasting the trends in the NASDAQ, a pivotal task for which it has not been explicitly trained, traditionally referred to as zero-shot learning. Instead of leveraging finance-specific data, we hinge on ChatGPT's intrinsic NLP capabilities. Moreover, we introduce an innovative prompting method to leverage ChatGPT's analytical processes by finding headlines, then converting these headlines into a sentiment, and finally aggregating carefully these scores with both a cumulated sum and a detrended process to filter out noise. Such insights not only augment the nascent literature on deciphering intricate news with LLM models but also differentiate our study from contemporaneous works that use chatGPT in a more brute-force way.

3. Prompt engineering

3.1. Data collection

We collected Bloomberg Global Markets Wrap summaries from 2010 to October 2023. We ignored any text that is less than 600 characters long or any news summary that is not explicitly a market wrap by removing any text that does not contain the keywords "market(s) wrap". Over 3600 news items were collected for applying a two-step approach detailed in section 3.2. Considering that these summaries encapsulate daily market developments across 10 to 20 headlines, the aggregate dataset is indicative of 36 to 72 thousand comprehensive news items, meticulously curated and verified.

3.2. Two-step approach

We opted to decompose the instructions into simpler and more straightforward tasks. In accordance with the recommendations posited in (Lopez-Lira and Tang, 2023), we devised two prompts to refine the objectives for ChatGPT, focusing on tasks empirically demonstrated to align well with ChatGPT's capabilities. Our first prompt consisted of summarizing the text into titles or headlines as follows:

First Prompt:

> *Assume you are an experienced asset manager. Analyze the text between {} and identify the predominant themes. For each theme, formulate a compelling headline that encapsulates its core message. Please arrange your responses in a list format, ensuring a line break after each headline.*
> *Your list should contain a total of 15 distinct headlines reflecting the respective themes and presented in the following format:*

1. Headline that encapsulates Theme 1
2. Headline that encapsulates Theme 2
...
15. Headline that encapsulates Theme 15
{INSERT_TEXT_HERE}

Our second prompt consisted of determining a sentiment score on each headline:
Second Prompt:

Assume you are an experienced asset manager. Your task is to assess the impact of various economic events and trends on global equities. For each numbered statement provided below between{}, classify its impact as either "positive," "negative," or "indecisive". {INSERT_TEXT_HERE}

For the two prompts, we used the gpt-4.0 version of ChatGPT. The overall idea of this two-step approach is to ease the task of chatGPT and leverage its capacity to make summaries and in a second step find the tone or sentiment. We can now devise an enhanced and more pertinent "Global Equities Sentiment Indicator".

4. Global Equities Sentiment Indicator

Definition 4.1. ***Daily Sentiment Score:*** *Let us denote h_i as the i^{th} headline scanned from the daily news n and have two scoring functions that are consistent, a positive one $p(h_i)$ which returns 1 if h_i is positive, 0 otherwise and a negative one $n(h_i)$ which returns 1 if h_i is negative, 0 otherwise.*
The sentiment score S for a day with N headlines is given by:

$$S = \frac{\sum_{i=1}^{N} p(h_i) - \sum_{i=1}^{N} n(h_i)}{\sum_{i=1}^{N} p(h_i) + \sum_{i=1}^{N} n(h_i)} \quad (1)$$

The sentiment score S measures the relative dominance of positive versus negative sentiments in a day's headlines. It satisfies a couple of simple properties that are trivial to prove. As described in table 1, once we have the daily individual positive and negative score, the sentiment score is easily computed. Moreover, the sentiment score satisfies some properties as highlighted in proposition 1.

Proposition 1. *The sentiment score S satisfies some properties:*

1. ***Boundedness**: S is bounded as $-1 \leq S \leq 1$.*
2. ***Symmetry**: If sentiments of all headlines are reversed, then S changes its sign.*
3. ***Neutrality**: $S = 0$ if there are equal numbers of positive and negative headlines.*
4. ***Monotonicity**: S increases as the difference between positive and negative headlines increases.*
5. ***Scale Invariance**: S remains the same if we multiply the number of both positive and negative headlines by a constant.*
6. ***Additivity**: The combined S for two sets of headlines is the weighted average of their individual S values.*

Date	Positive	Negative	Score
2010-01-04	11	3	0.57
2010-01-05	6	6	0.00
...			
2023-11-21	8	3	0.45

Table 1: Sentiment Analysis Dataset

Figure 1 depicts the raw signal corresponding to the score, which exhibits significant noise. Using raw sentiment scores from daily news headlines often results in noisy and less interpretable outcomes. To address this, we propose a *cumulated sentiment score* over a specified period. This score aggregates news sentiments over a duration, offering a more comprehensive measure of the news impact during that period.

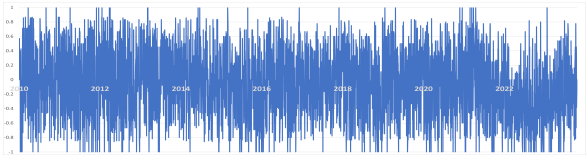

Figure 1: Raw signal exhibiting significant noise

Definition 4.2. ***Cumulated Sentiment Score:****We defined a cumulative score as follows. Given:*

- *$h_{i,t}$ as the i^{th} headline on day t.*
- *$p(h_{i,t})$ and $n(h_{i,t})$ as functions returning 1 for positive and negative sentiments of $h_{i,t}$ respectively, 0 otherwise.*
- *d as the duration.*

The cumulated sentiment score S_d over period d is:

$$S_d = \frac{\sum_{t=1}^{d} \sum_{i=1}^{N_t} p(h_{i,t}) - \sum_{t=1}^{d} \sum_{i=1}^{N_t} n(h_{i,t})}{\sum_{t=1}^{d} \sum_{i=1}^{N_t} p(h_{i,t}) + \sum_{t=1}^{d} \sum_{i=1}^{N_t} n(h_{i,t})} \quad (2)$$

with N_t being the number of headlines on day t.

The mathematical properties of proposition 1, that is boundedness, symmetry, neutrality, monotonicity, scale invariance remains for the

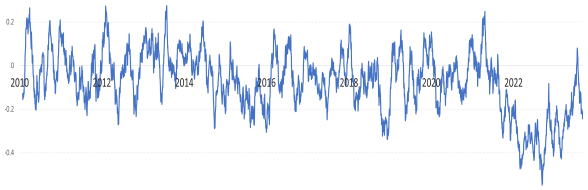

Figure 2: Cumulated sentiment score with d=20

cumulated sentiment score. Figure 2 illustrates how the cumulated process diminishes the noise within the signal.

The cumulative sentiment enabled us to obtain the trend of the news rather than a momentary snapshot of it, which appeared to be informative.

5. Evaluation of the Sentiment Score's Validity

5.1. Descriptive statistics

In order to evaluate the performance of our sentiment score to reveal information about the market reaction, we consider two correlation metrics: Pearson and Spearman coefficient as presented in (Wilcox, 2010). While Pearson correlation coefficients capture linear relationship, the Spearman rank correlation coefficients are a measure of the monotonic relation between the two variables thanks to the ordering of the rank functions and can deal with ordinal or non-normally distributed data, providing a robust measure of association for non linear data.

5.2. The Equity Data and Variable Computation

To assess the robustness of the score, we computed its correlation with diverse equity markets: the SP 500, NASDAQ 100, Nikkei 225, Eurostoxx 50, FTSE 100, and MSCI Emerging Countries indices. We call these markets respectively US, US Tech, Japan, Europe, UK and Emerging equities markets or simply by their region without mentioning equities market explicitly. We used data from January 2010 to November 2023 and computed the resulting returns over multiple periods $(p_i)_{i=1..n}$ to measure the horizon for which the sentiment score is predictive as follows:

$$R_{t+1}^{p_i} = \frac{P_t - P_{t-p_i}}{P_{t-p_i}}$$

- $R_{t+1}^{p_i}$: The return over the p_i period of the equity at time $t+1$.
- P_t: The value of the equity at current time t.
- P_{t-p_i}: The value of the equity at a p_i period before the current time.

On purpose, the return $R_{t+1}^{p_i}$ is time stamped at time $t+1$ to avoid any data leakage and ensures that we have all the relevant data at the time of the computation.

5.3. Correlation Results

The aim is to measure the correlation between future equity market returns and the cumulative sentiment score calculated over different periods. Hence, we computed both Pearson and Spearman coefficients to evaluate the relationship between these variables two-by-two. The correlation matrices are of size 49 by 49, hence contain 2401 elements.

The first experiment was to validate the difference in correlation provided by different periods for the cumulative sentiment score and forward returns. We provide in figure 3 the result for the US Tech market. Figures 8, 9, 10, 11, 12 provide the results for the other markets, namely US, Japan, Europe, UK and Emerging markets for the Pearson correlation matrices. Likewise, figures 25, 26, 27, 28, 29, 30 provide the Spearmann correlation matrices for the same markets.

The overall correlation between sentiment scores and future returns is positive, as evidenced by the predominantly red color of the matrices. This positive correlation tends to increase with longer periods for both cumulative sentiment scores and forward returns, forming a diagonal pattern. However, for very long period of future returns we observe a negative correlation.

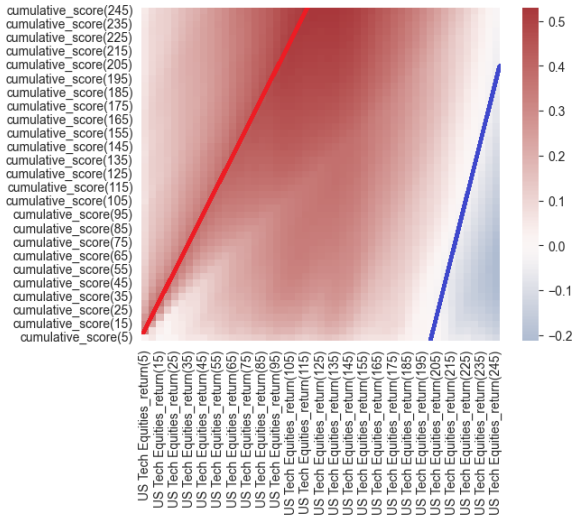

Figure 3: Pearson correlation matrix of the cumulative score and the NASDAQ

These results are consistent across markets, suggesting that the approach is robust and generalizable.

Figure 3 showcases a red diagonal highlighting the presence of positive correlation values, while a blue diagonal signifies negative correlation patterns. In the case of positive correlation, a diagonal composed of the highest values is surrounded by other elevated values, with the values diminishing as they move away from the diagonal. We observe that the values increase for longer periods of the cumulated sentiment score. Moreover, the negative correlation pattern is evident in the long-term market return, characterized by a diagonal of non-correlated values, with a decrease of these values observed to the right of this diagonal. This pattern exists in all the other markets as proved in section 7.

5.4. T-test on the correlation

In order to validate the statistical significance of the correlation values, we applied a t-test to all the results. We focused on the p-value associated to each test. Because the number of conducted test is very large, we do all our T-test using the False Discovery Rate method.

5.4.1. False Discovery Rate

The False Discovery Rate (FDR) is a statistical method crucial for managing the challenge of multiple comparisons in large-scale experiments, as introduced by (Benjamini and Yekutieli, 2001). In contexts where numerous statistical tests are conducted simultaneously, the FDR addresses the increased risk of false positives by controlling the expected proportion of false discoveries among all significant results. This approach effectively regulates the false selection rate, ensuring that only a predetermined percentage of rejected hypotheses are likely to be false positives. The procedure is employed to rank p-values and determine a critical threshold, enabling to identify statistically significant results while managing the trade-off between sensitivity and specificity.

5.4.2. T-Test Adaptation

In a two-tailed t-test, the p-value signifies the probability of observing a t-statistic as extreme as the one calculated from the sample data, assuming the null hypothesis holds. For correlation values, the null hypothesis typically posits no significant correlation between the variables. The FDR adapts the threshold for improving the statistical significance assessment in a large experiment case.

In figure 4 we plot in white all the correlations whose p-value FDR adapted are below one percent and the rest, that is to say, the correlation values where we fail to reject the null hypothesis of a non significant correlation value in grey with a color scale. Most of the correlation matrix is white indicating that the correlation numbers are mostly statistically significant. Like what we did for the correlation analysis, we can validate the tests on other equity markets. Figures 13, 14, 15, 16, 17, 18 show that all equitites markets exhibit similar behavior for the p-values of Pearson correlation while figures 31, 32, 33, 16, 17, 36 show that all equitites markets exhibit similar behavior for the p-values of Spearman correlation

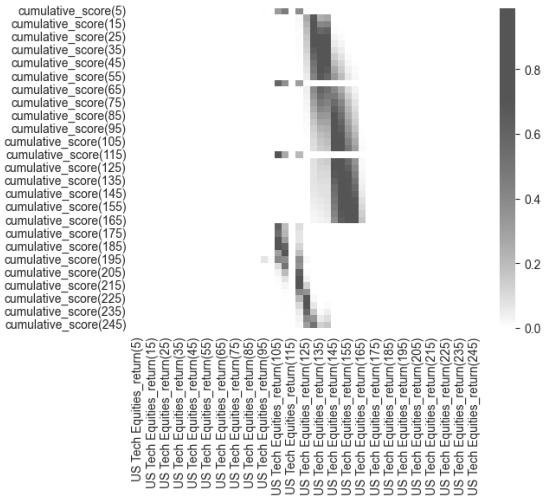

Figure 4: Adjusted p-value for the Pearson correlation between the US Tech market and the cumulative sentiment score

5.4.3. The Mitigated Matrix

Consideration should be given exclusively to correlation values demonstrating statistical significance. Our aim is to adjust each correlation in accordance with its corresponding p-value. As illustrated in Figure 4, the correlation matrix is modified using a gradient approach. Specifically, correlation values are retained as-is when p-values suggest statistical significance. Conversely, in instances of increasing p-value, the correlations are adjusted as follows:

$$\rho_{i,j}^{\text{mitigated}} = \rho_{i,j} \times (1 - p_{i,j}) \tag{3}$$

Here, $\rho_{i,j}$ represents the correlation coefficient, and $p_{i,j}$ denotes the associated p-value. Figure 5 displays the resulting mitigated correlation matrix. This method allows for the prioritization of statistically significant correlations without excessive discrimination.

Analysis reveals that the matrix's region of interest is predominantly significant. Non-significant values are found in longer horizons for cumulative_score and Equity return. These findings are

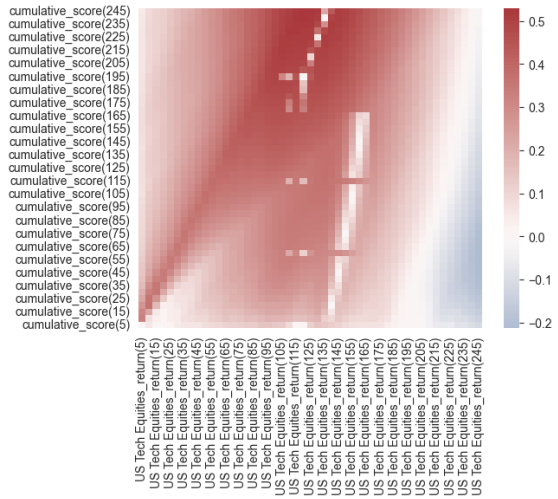

Figure 5: Mitigated correlation between the US Tech market and the cumulative sentiment score

applicable across various equity markets for both Pearson and Spearman correlations. Figure 19, 20, 21, 23, 22, 24, 37, 38, 39, 41, 40, 42 corroborate these results.

5.4.4. The Short Term Correlation

The correlation exhibits notable values within the range of $[-0.30, 0.53]$. To emphasize statistical significance, we focus exclusively on the matrix section where p-values are below 0.01, representing the white area. This selection yields significant correlation results for the mid-term return and cumulative sentiment score lag.

5.4.5. The Best Combinations

Among all the coefficients, we can exclude the non significant ones according to a t-test, hence with p-value exceeding the 0.01 threshold. We could obtain the duo of variables that obtain the highest and lowest correlation values in table 2 and 3 respectively.

Score	Equity	Positive Correlation
S_{245}	US Tech(125)	0.53
S_{205}	US(95)	0.47
S_{245}	Japan(135)	0.27
S_{80}	Europe(35)	0.22
S_{80}	UK(35)	0.25
S_{245}	Emerging(125)	0.43

Table 2: Highest Pearson positive correlation values by equities

We remind that S_d represents the cumulative score denoted in the matrix as "cumulative_score(d)".

Score	Equity	Negative Correlation
S_{245}	US Tech(245)	-0.26
S_5	US(245)	-0.31
S_{245}	Japan(245)	-0.19
S_5	Europe(245)	-0.30
S_5	UK(210)	-0.18
S_{245}	Emerging(245)	-0.14

Table 3: Highest Pearson negative correlation values by equities

The analysis reveals a clear, positive relationship between the cumulative score and equity returns, with the strength of the correlation intensifying as the lag size of the cumulative score increases. Interestingly, as we delve into deeper cumulative scores, the negative correlation diminishes. There is a discernible trade-off concerning the lag of the cumulative score: seeking an optimal balance is crucial, as the cumulative score lags behind the equity market. We aim to maximize the correlation while maintaining a current score reflective of the market's status. For instance, opting for a substantial lag in the cumulative score may yield a strong correlation, yet the estimator's time relevance could be compromised. This dynamic is evident in the correlation matrix, where red signifies positive correlation and blue indicates negative correlation, guiding us towards a precise analysis. Markets demonstrate different degrees of sensitivity to the timing of news, with the cumulative score's correlation extending over a more extended period than previously observed with sentiment scores. The investigation into the relationship between cumulative scores and equity returns illuminates the crucial dynamics of lag impact. The subsequent section will delve into the intricate trade-off that exists between the lag value of the score and the intensity of the signal it provides.

6. Trade-Off Analysis of Financial Indicators

The investigation into the relationship between cumulative scores S_d and equity returns unveils the pivotal dynamics of market reaction delays. The forthcoming analysis explores the nuanced trade-off between the depth of the cumulative score—reflected by the subscript d in S_d—and the predictive signal's intensity it conveys. The term d represents the depth of analysis, encapsulating the cumulative effect of sentiment over a defined period.

The depth of the cumulative score, denoted as S_d, is mathematically defined as the aggregate sentiment measured over a period d. This period reflects the span over which the sentiment data is cumulated, not to be confused with the delay in

market reaction. The delay in market impact is instead associated with the temporal shift applied to the equity return data, which is examined against the cumulative sentiment scores.

The correlation value, represented by ρ, quantifies the strength and direction of the linear relationship between the financial indicator's cumulative score S_d and the shifted equity returns. The mean correlation value for different prediction horizons, ranging from 1 to 12 months, is computed as follows:

$$\bar{\rho}_{\text{horizon}}(i) = \frac{1}{s \times (j+1)} \sum_{k=1}^{s\times(j+1)} \rho_{i,k} \qquad (4)$$

where $\bar{\rho}_{\text{horizon}}(i)$ is the mean correlation at the i^{th} cumulative value for a given horizon, and $\rho_{i,k}$ is the correlation value at the i^{th} cumulative value for the k^{th} shifted time point within the horizon. The term $s \times (j+1)$ denotes the number of discrete time intervals encapsulated within the horizon, where $j \in \{0, 1, \ldots, 11\}$ and s is the number of equity return included for mean computation.

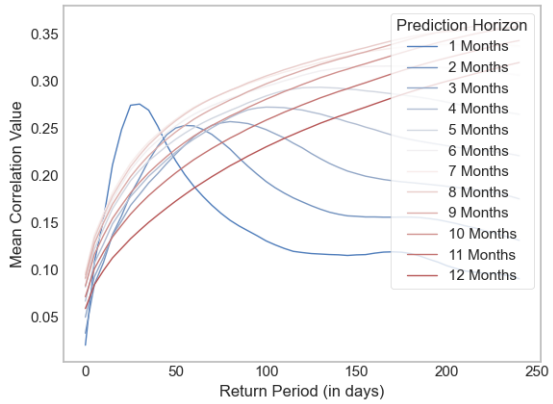

Figure 6: US Tech Equity: Mean correlation of cumulative score against shifted returns across horizons

The objective of the analysis is to determine the optimal depth d_{opt} of S_d that maximizes ρ, while still being timely enough to provide practical predictive utility for market reactions. This optimal point is characterized by the highest mean correlation value that can be achieved before the utility of the cumulative score is compromised by its stale reflection of market sentiment. Like what we did for the correlation analysis, we can perform the same analysis for the other equity markets. This is provided by figures 43, 44, 45, 46, 47, 48.

6.1. Optimal Point Determination

The apex of the curve in Figure 6 indicates the optimal depth d_{opt} of S_d, at which the mean correlation $\bar{\rho}$ is maximized. This peak represents the ideal balance between comprehensive sentiment analysis and timely market prediction, ensuring the cumulative score's relevance and predictive power.

To ascertain d_{opt}, we locate the curve's highest point, which signifies the strongest linear relationship between S_d and market performance, without undue delay. Table 4 provides the optimal period for each equity market over a prediction horizon of one month. The step size s is 4 and includes the market return on 20 days.

Equity	d_{opt}
US Tech	40
US	30
Japan	40
EU	25
UK	30
EM	30

Table 4: Mean correlation values for different equities over one month

7. Robustness over the Equities Markets

This section examines the robustness of the identified pattern across different equity markets. The question arises whether a universal pattern exists within these markets. To address this, we compare each matrix with the average correlation matrix, representing the common pattern, and assess the distance in terms of standard deviation of each matrix from this common pattern. Like for the rest of the paper on other indicators, we can notice consistency across equities markets.

7.1. Computation of Mean Matrix and Standard Deviation

The mean matrix, denoted as Z, is computed as the average of all correlation matrices:

$$Z = \frac{1}{n}\left(\sum_{k=1}^{n} m_{i,j}^{k}\right) \qquad (5)$$

where $\left(m_{i,j}^{k}\right)$ represents the i,j correlation matrix coefficient of the k market and n is the total number of markets. Strictly speaking, the mean matrix is computed for each cell as the mean across all markets. Likewise, for each matrix cell, we compute the standard deviation of correlations across all markets

$$\Sigma(Z) = \frac{1}{\sqrt{n-1}}\left(\sum_{k=1}^{n}\left(m_{i,j}^{k} - z_{i,j}\right)^2\right)$$

where $z_{i,j} = \sum_{k=1}^{n} m_{i,j}^{k}/n$ is the coefficient of the mean matrix presented in equation 5

7.2. Element-wise T-test Analysis

In order to ensure proprer resizing of each correlation as well as the average correlation matrix, we first z-score them as follows:

$$\tilde{M}_{ij}^{k} = \frac{M_{ij}^{k} - \bar{M}_{ij}^{k}}{\Sigma(M)_{ij}} \tag{6}$$

For each market matrix with upper index k, we conduct an element-wise T-test comparing it to the mean matrix Z. The T-statistic is computed elementwise as:

$$T_{ij} = \frac{M_{ij}^{k} - Z_{ij}}{\Sigma(Z)_{ij}} \tag{7}$$

The p-values are computed using two-tails test:

$$p = 1 - 2 \times (1 - \text{CDF}_{student}(|T|)) \tag{8}$$

7.3. Analysis of P-Value Results

Table 5 presents the percentage of each equity market matrix where the p-value falls below the 0.01 significance threshold:

Equity Market	% of Matrix
US Tech	80
US	91
Japan	92
Euro	86
United Kingdom	55
Emerging	75

Table 5: Proportion of Each Equity Matrix Validating the Common Pattern

A score of 100% implies that the matrix perfectly follows the pattern of the mean matrix, while a score of 0% indicates no common pattern with the mean matrix.

The results indicate a significant presence of the identified pattern across all markets, with an especially pronounced effect in the Japanese market (99%). The US Technology and US General markets exhibit substantial percentages (78% and 69% respectively). This variation suggests a differential impact of sentiment scores on equity returns across these markets.

The high percentages in the Euro, United Kingdom, and Emerging Markets (ranging from 84% to 94%) further reinforce the ubiquity of the pattern. These findings collectively suggest that sentiment scores consistently influence equity returns across diverse global markets, underpinning the robustness of the identified pattern.

This analysis confirms the existence of a common pattern across various equity markets, linking sentiment scores to equity returns. The consistency of significant p-values across markets underscores the widespread impact of investor sentiment on market movements, presenting valuable insights for market analysis and investment strategies.

7.4. Matrix quantile distance

A second method consists in doing a quantile difference test between each market correlation matrix and the average over each market. Although this approach is less well-known than the standard correlation t-test, converting correlation matrices into quantiles for each cell and then computing their average absolute difference to judge the quantile distance is a method to judge if two matrices share a similar profile. This approach makes sense for several reasons:

- **Robustness:** Quantiles are less affected by outliers compared to raw correlation values. This can give a more robust comparison, especially in the presence of extreme values.
- **Normalization:** It normalizes the scale of comparison. Since correlation coefficients are bounded between -1 and 1, converting them to quantiles puts them on a uniform scale.
- **Sensitivity to Distribution:** This method is sensitive to the distribution of correlation coefficients across the matrices. By using quantiles, you're comparing the relative positions of correlation coefficients, which can be more informative about the similarity in patterns of correlation.
- **Interpretable Metric:** The average absolute difference is an easily interpretable metric that quantifies the average discrepancy between the matrices in terms of their quantile-transformed correlations.

Mathematically, if C_1 and C_2 are two correlation matrices, converting them to quantiles involves replacing each correlation coefficient with its corresponding quantile rank within the matrix that we denote for each matrix i, j cell as Q_{ij}^1 and Q_{ij}^2 respectively. The average absolute difference is calculated as $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Q_{ij}^1 - Q_{ij}^2|$, where n is the dimension of the matrices. This value gives an overall measure of how different the two matrices are in their correlation structure.

Table 6 displays the proportion of each equity market matrix with quantile difference above ten percents. The results exhibit consistency with the previous method table 5, confirming us that the sentiment news is consistent accross major equities markets.

Equity Market	% of Matrix
US Tech	80
US	99
Japan	92
Euro	89
United Kingdom	51
Emerging	72

Table 6: Proportion of Each Equity Matrix Validating the Common Pattern using quantile distance over 10%

8. Conclusion

In this paper, we look at the equity market reaction to market news sentiment. We document significant correlations between news market sentiment and equity returns regarding the cumulative sentiment score. We also show that the correlation reverts to a negative correlation over longer horizons. We validate that this behavior exists in other equity markets, validating the robustness of the pattern. We suggest an optimal period that balances the trade-off between the market's reactivity to new information and the strength of correlation between sentiment score and forward equities returns.
Future research could elaborate on this sentiment score to suggest a systematic NLP based long short strategy on world wide equity indices.

9. Bibliographical References

Douglas W Arner, Janos Barberis, and Ross P Buckley. 2015. The evolution of Fintech: A new post-crisis paradigm. *Geo. J. Int'l L.* 47 (2015), 1271.

Y. Benjamini and D. Yekutieli. 2001. Control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29, 1 (2001), 1165–1188.

Tyler Cowen and Alexander T. Tabarrok. 2023. How to Learn and Teach Economics with Large Language Models, Including GPT. *SSRN Electronic Journal* XXX, XXX (3 2023), 0–0. https://doi.org/10.2139/SSRN.4391863

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* XX, XX (2018), XX.

Georgios Fatouros, Georgios Makridis, Dimitrios Kotios, John Soldatos, Michael Filippakis, and Dimosthenis Kyriazis. 2023. DeepVaR: a framework for portfolio risk assessment leveraging probabilistic deep neural networks. *Digital finance* 5, 1 (2023), 29–56.

A Shaji George and AS Hovan George. 2023. A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal* 1, 1 (2023), 9–23.

Abbas Ghaddar and Philippe Langlais. 2020. Sedar: a large scale French-english financial domain parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 3595–3602. http://www.lrec-conf.org/proceedings/lrec2020/index.html

Anne Lundgaard Hansen and Sophia Kazinnik. 2023. Can ChatGPT Decipher Fedspeak? *SSRN Electronic Journal* XX, XX (3 2023), XX. https://doi.org/10.2139/SSRN.4399406

Ioan-Bogdan Iordache, Ana Sabina Uban, Catalin Stoean, and Liviu P Dinu. 2022. Investigating the Relationship Between Romanian Financial News and Closing Prices from the Bucharest Stock Exchange. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 5130–5136. http://www.lrec-conf.org/proceedings/lrec2022/index.html

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 2293–2299. http://www.lrec-conf.org/proceedings/lrec2020/index.html

Hyungjin Ko and Jaewook Lee. 2023. Can Chatgpt Improve Investment Decision? From a Portfolio Management Perspective. *SSRN Electronic Journal* XX, XX (2023), XX. https://doi.org/10.2139/SSRN.4390529

Anton Korinek. 2023. Language Models and Cognitive Automation for Economic Research. *Cambridge, MA* XX, XX (2 2023), XX. https://doi.org/10.3386/W30957

Chenying Li, Wenbo Ye, and Yilun Zhao. 2022. Finmath: Injecting a tree-structured solver for question answering over financial reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 6147–6152. http://www.lrec-conf.org/proceedings/lrec2022/index.html

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. ICLR, ICLR, 4513–4519.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Electronic Journal* XXX, XX-XX (4 2023), XX. https://doi.org/10.2139/SSRN.4412788

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance* 66, 1 (2011), 35–65.

Corentin Masson and Patrick Paroubek. 2020. NLP analytics in finance with DoRe: a French 250M tokens corpus of corporate annual reports. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 2261–2267. http://www.lrec-conf.org/proceedings/lrec2020/index.html

Antonio Moreno-Ortiz, Javier Fernández-Cruz, and Chantal Pérez Chantal Hernández. 2020. Design and evaluation of SentiEcon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference (LREC). LREC, LREC, 5065–5072. http://www.lrec-conf.org/proceedings/lrec2020/index.html

Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *SSRN Electronic Journal* XX, XX (3 2023), XX. https://doi.org/10.2139/SSRN.4375283

Joel Oksanen, Abhilash Majumder, Kumar Saunack, Francesca Toni, and Arun Dhondiyal. 2022. A Graph-Based Method for Unsupervised Knowledge Discovery from Financial Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 5412–5417. http://www.lrec-conf.org/proceedings/lrec2022/index.html

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* 37 (2017), 98–125.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108 (2016), 42–49.

Robert P Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)* 27, 2 (2009), 1–19.

Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62, 3 (6 2007), 1139–1168. https://doi.org/10.1111/J.1540-6261.2007.01232.X

Rand R. Wilcox. 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. Springer, NewYork.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023. The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges. *arXiv preprint arXiv:2304.05351* XX, XX (4 2023), XX.

Kai-Cheng Yang and Filippo Menczer. 2023. *Large language models can rate news outlet credibility*. Technical Report. arxiv. https://arxiv.org/abs/2304.00228v1

Chaofa Yuan, Yuhan Liu, Rongdi Yin, Jun Zhang, Qinling Zhu, Ruibin Mao, and Ruifeng Xu. 2020. Target-based sentiment annotation in Chinese financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 5040–5045. http://www.lrec-conf.org/proceedings/lrec2020/index.html

Thomas Yue, David Au, Chi Chung Au, and Kwan Yuen Iu. 2023. Democratizing financial knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology. *Available at SSRN 4346152* XX, XX (2023), XX.

Nadhem Zmandar, Tobias Daudert, Sina Ahmadi, Mahmoud El-Haj, and Paul Rayson. 2022. CoFiF Plus: A French Financial Narrative Summarisation Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. LREC, LREC, 1622–1639. http://www.lrec-conf.org/proceedings/lrec2022/index.html

A. Appendix

A.1. Cumulative Sentiment Score

A.1.1. Pearson Correlation Results

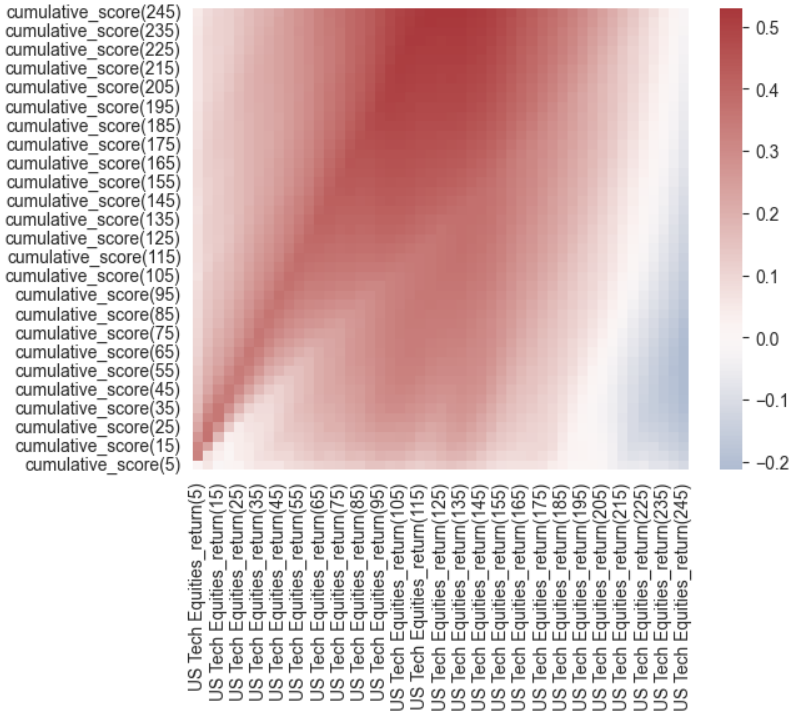

Figure 7: Pearson correlation between the USTech and the cumulative sentiment score

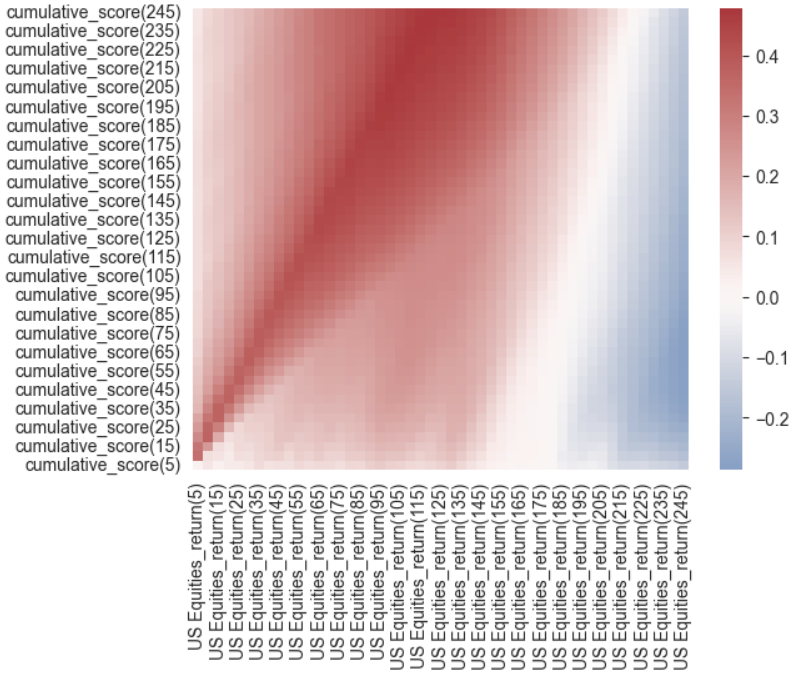

Figure 8: Pearson correlation between the US and the cumulative sentiment score

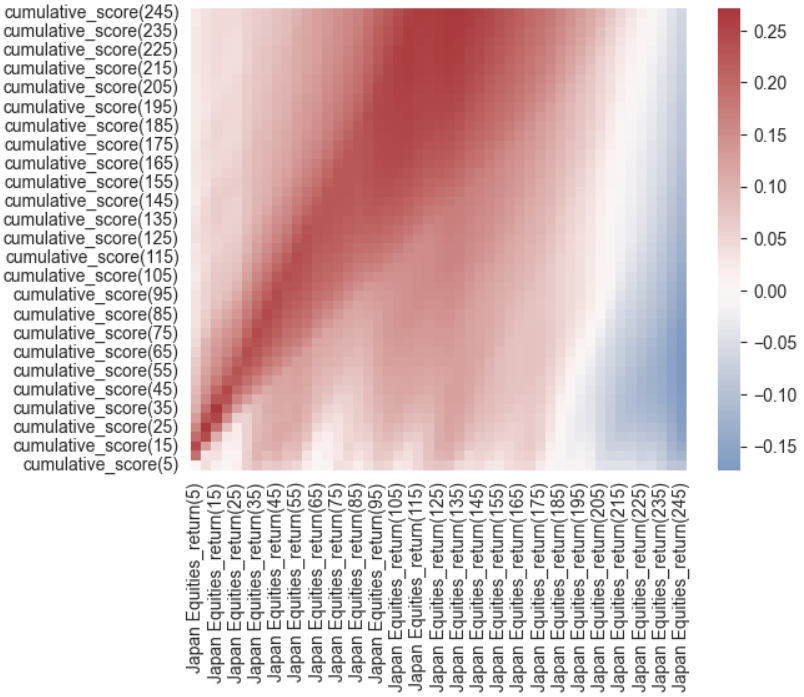

Figure 9: Pearson correlation between Japan and the cumulative sentiment score

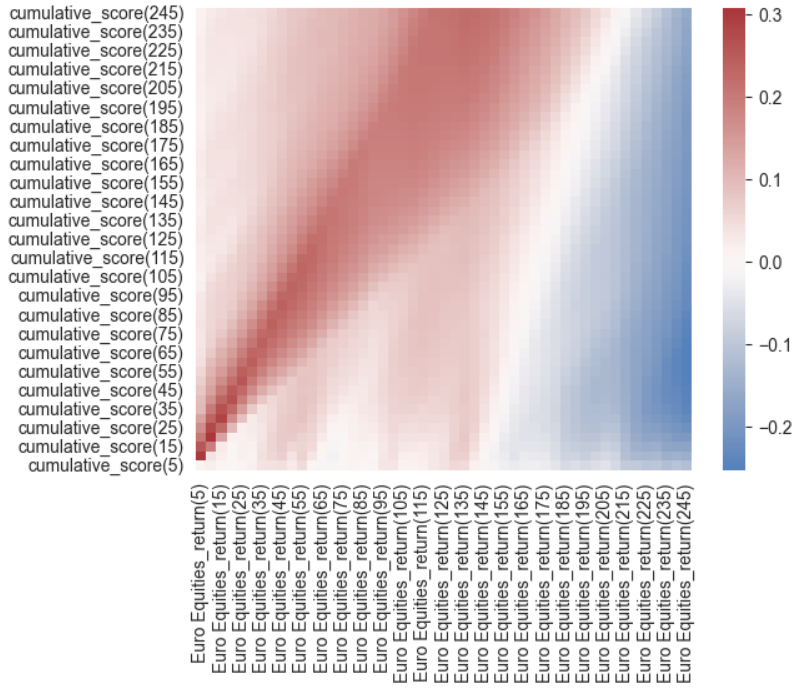

Figure 10: Pearson correlation between Euro and the cumulative sentiment score

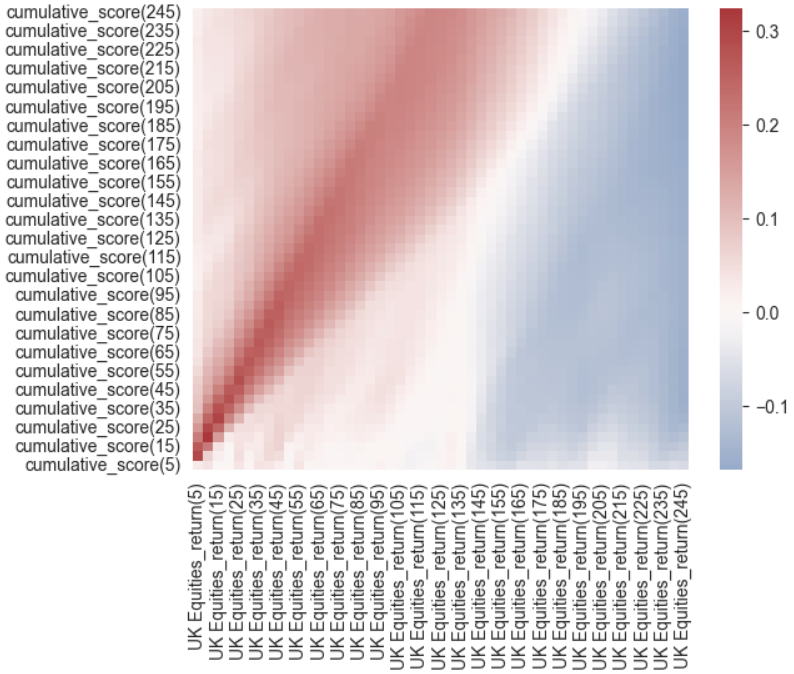

Figure 11: Pearson correlation between the UK and the cumulative sentiment score

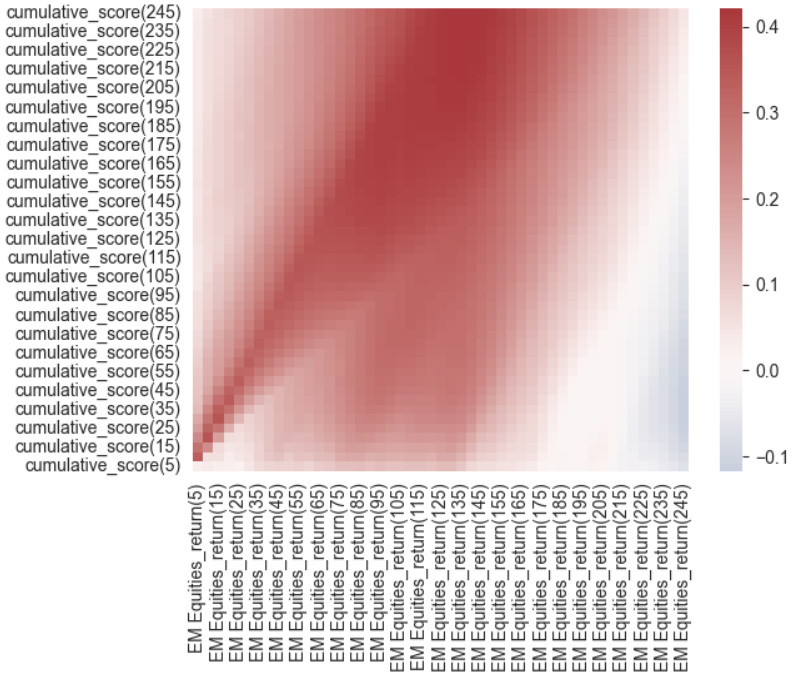

Figure 12: Pearson correlation between EM and the cumulative sentiment score

A.1.2. P-value Pearson Correlation Results

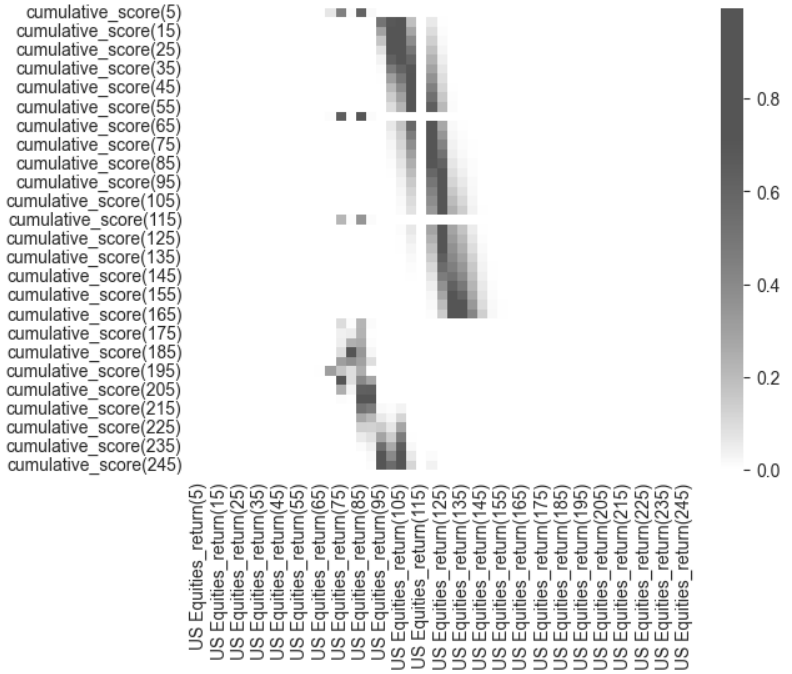

Figure 13: P-value for Pearson correlation between the U.S. and the cumulative sentiment score

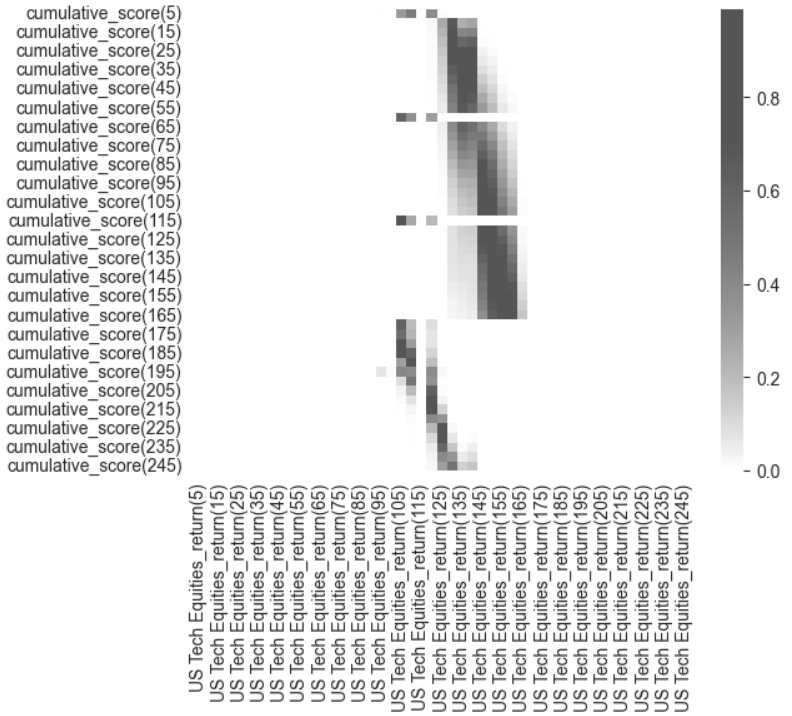

Figure 14: P-value for Pearson correlation between USTech and the cumulative sentiment score

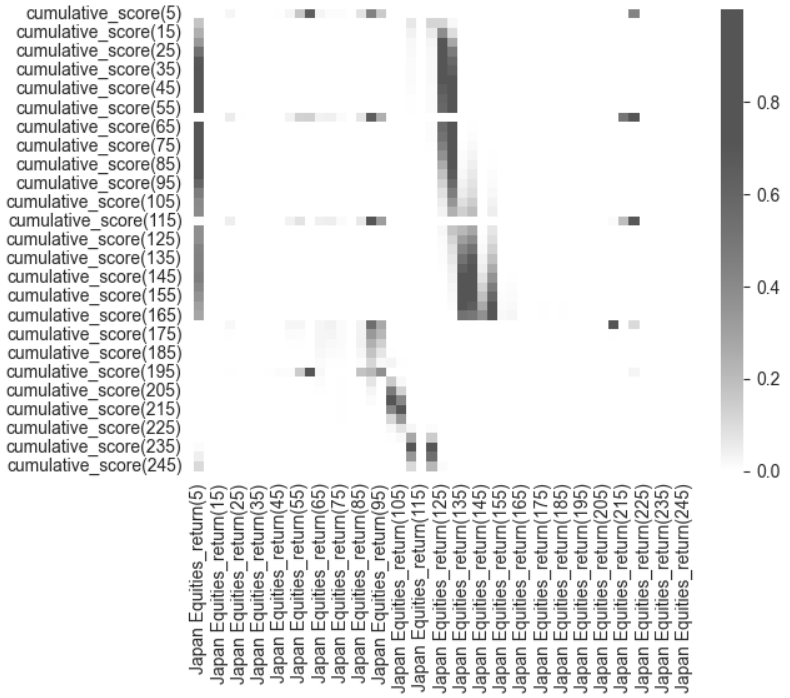

Figure 15: P-value for Pearson correlation between Japan and the cumulative sentiment score

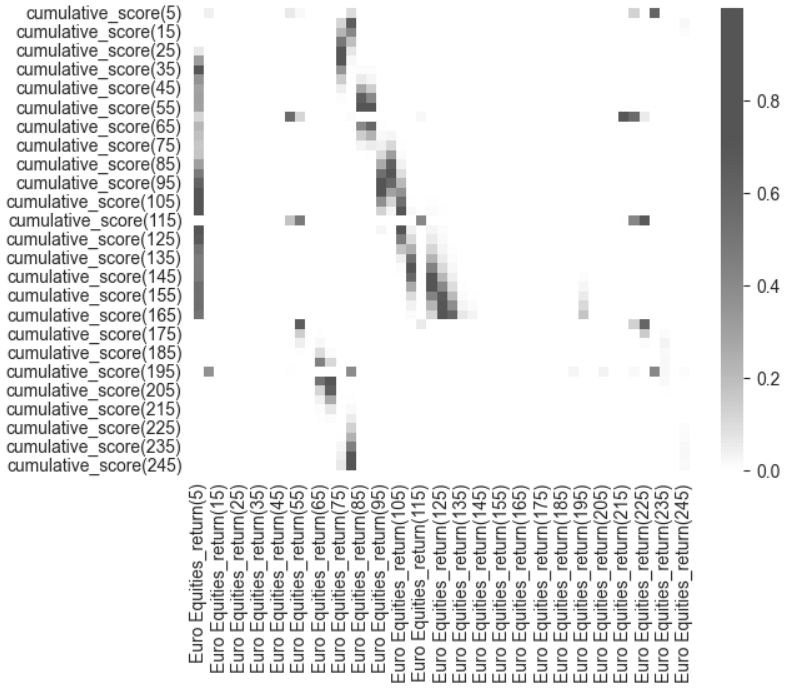

Figure 16: P-value for Pearson correlation between Euro and the cumulative sentiment score

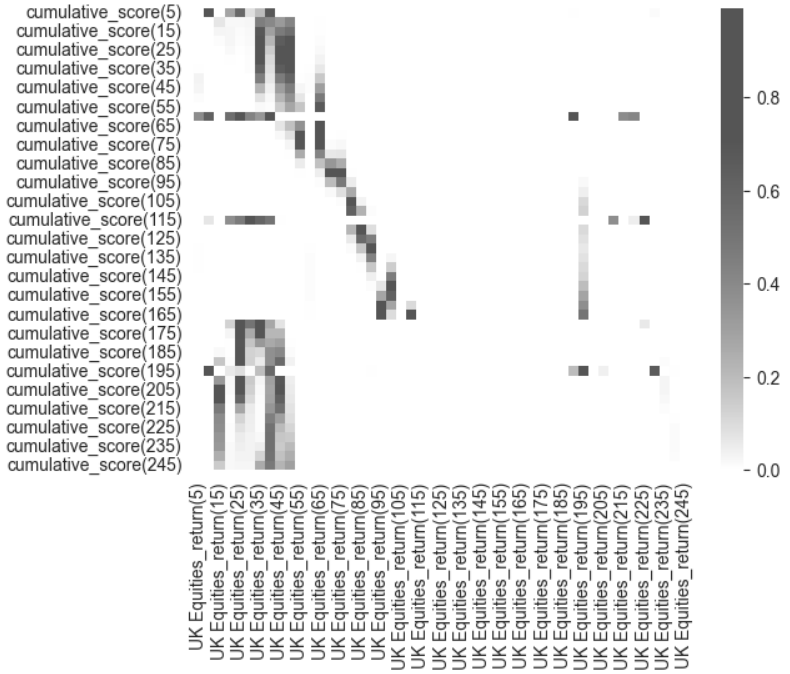

Figure 17: P-value for Pearson correlation between the UK and the cumulative sentiment score

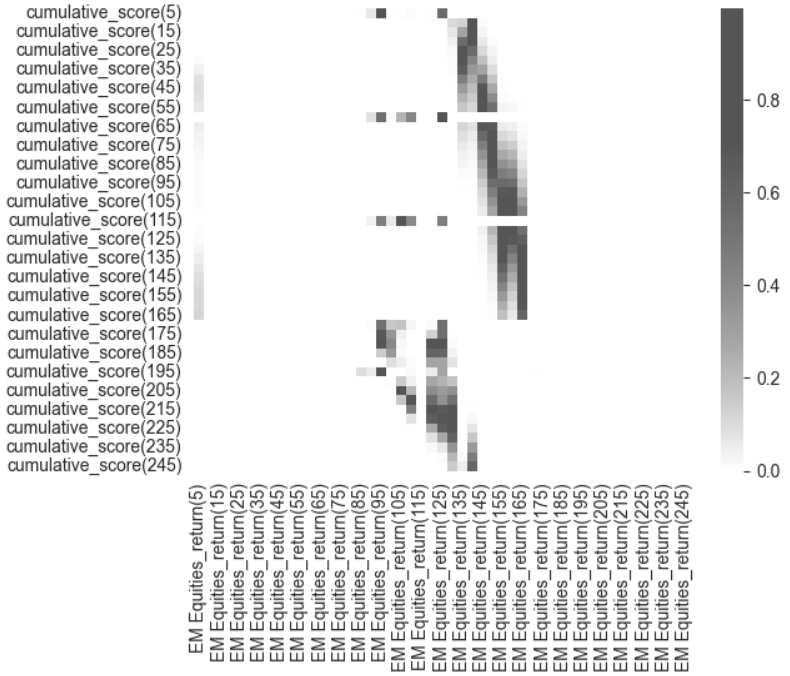

Figure 18: P-value for Pearson correlation between EM and the cumulative sentiment score

A.1.3. Mitigated Pearson Correlation Results

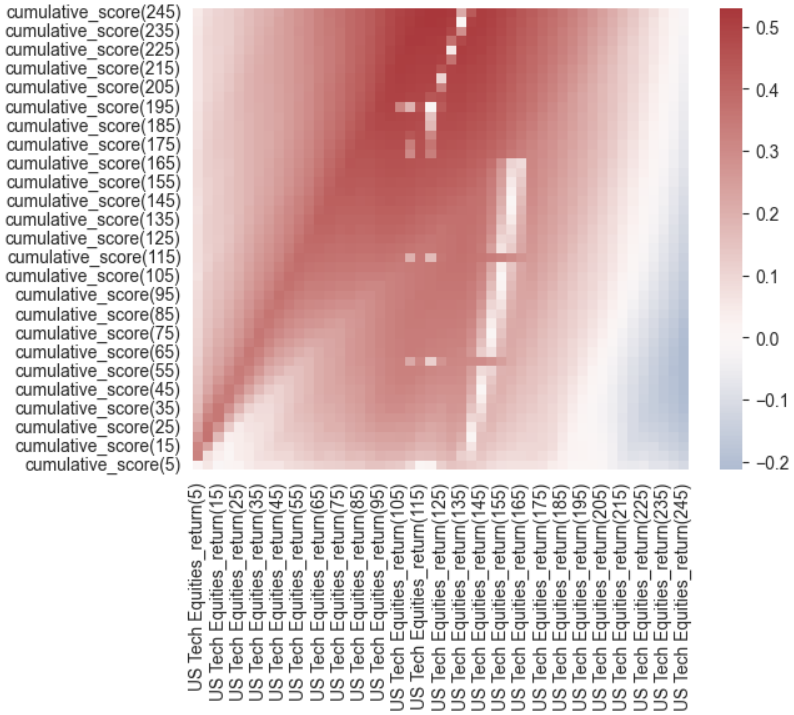

Figure 19: Mitigated Pearson correlation between the USTech and the cumulative sentiment score

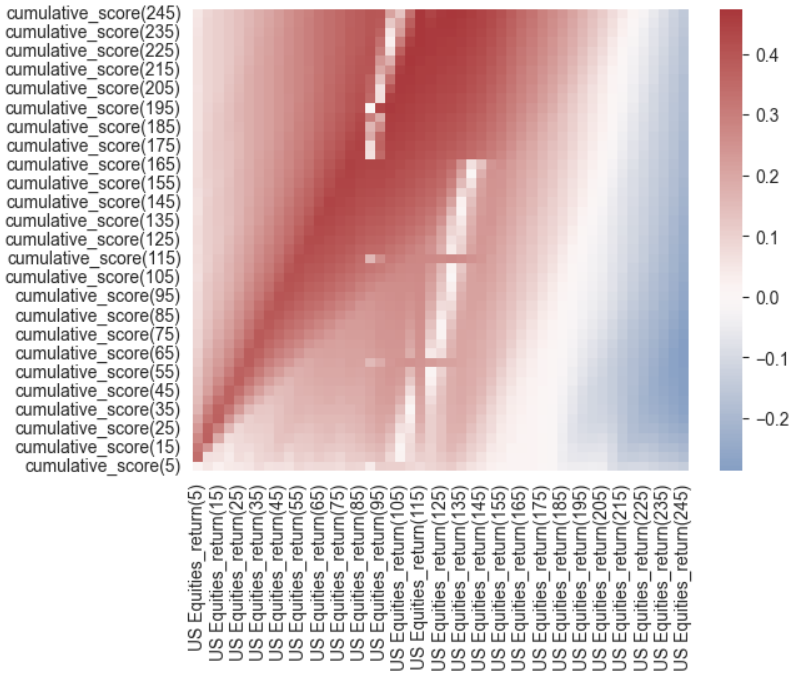

Figure 20: Mitigated Pearson correlation between the US and the cumulative sentiment score

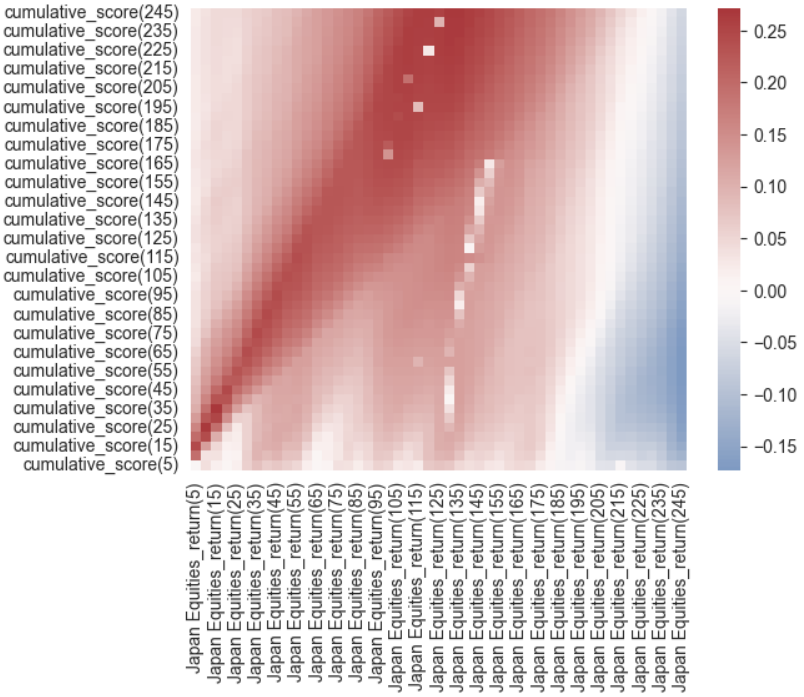

Figure 21: Mitigated Pearson correlation between Japan and the cumulative sentiment score

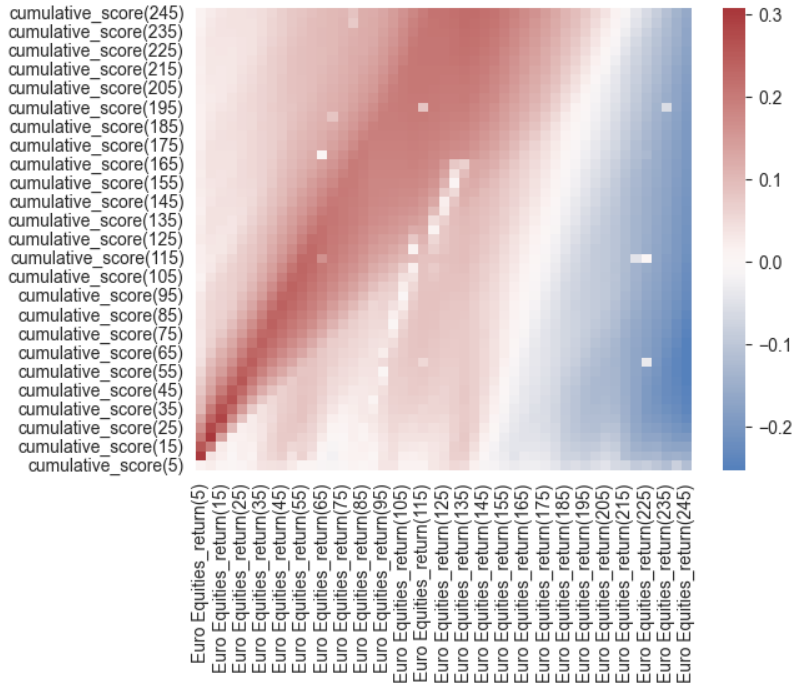

Figure 22: Mitigated Pearson correlation between Euro and the cumulative sentiment score

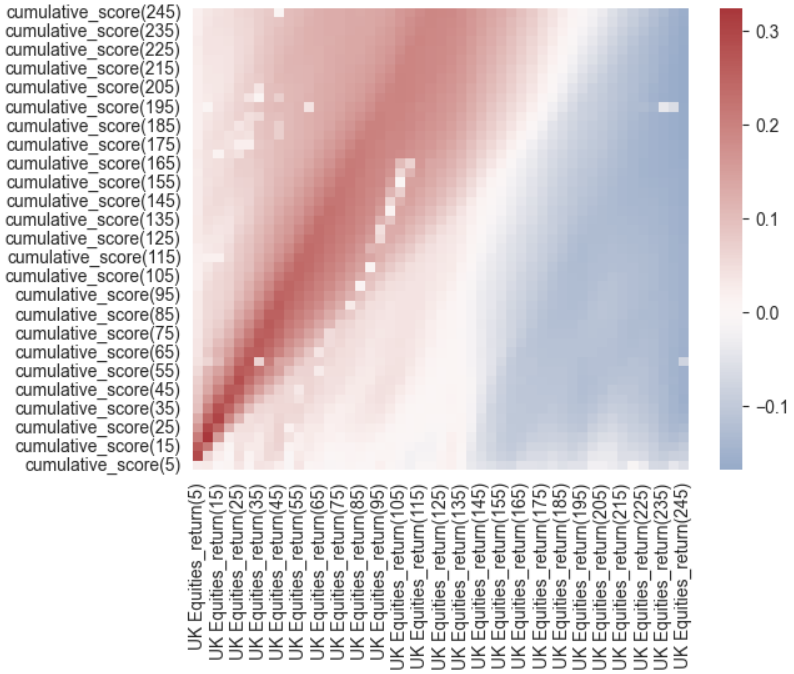

Figure 23: Mitigated Pearson correlation between the UK and the cumulative sentiment score

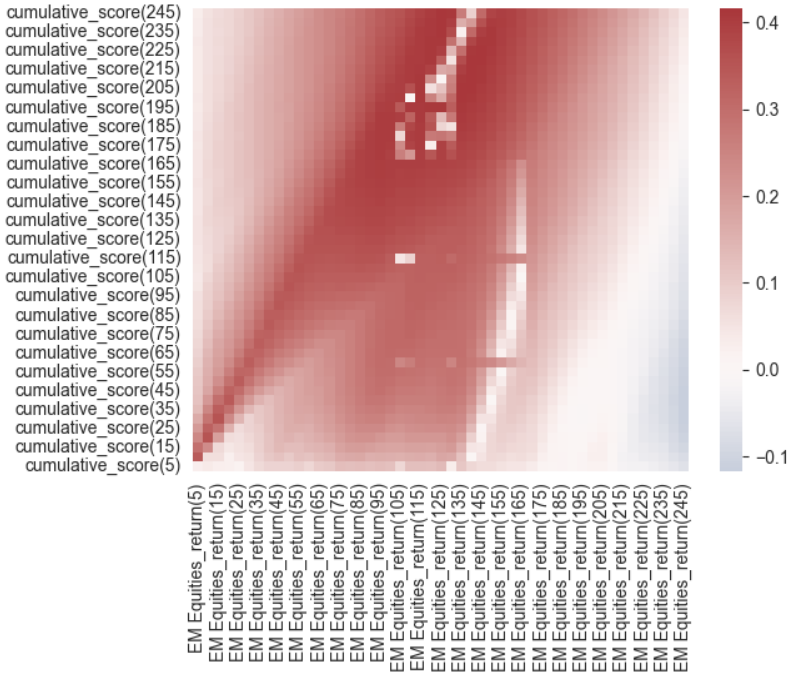

Figure 24: Mitigated Pearson correlation between EM and the cumulative sentiment score

A.1.4. Spearman Correlation Results

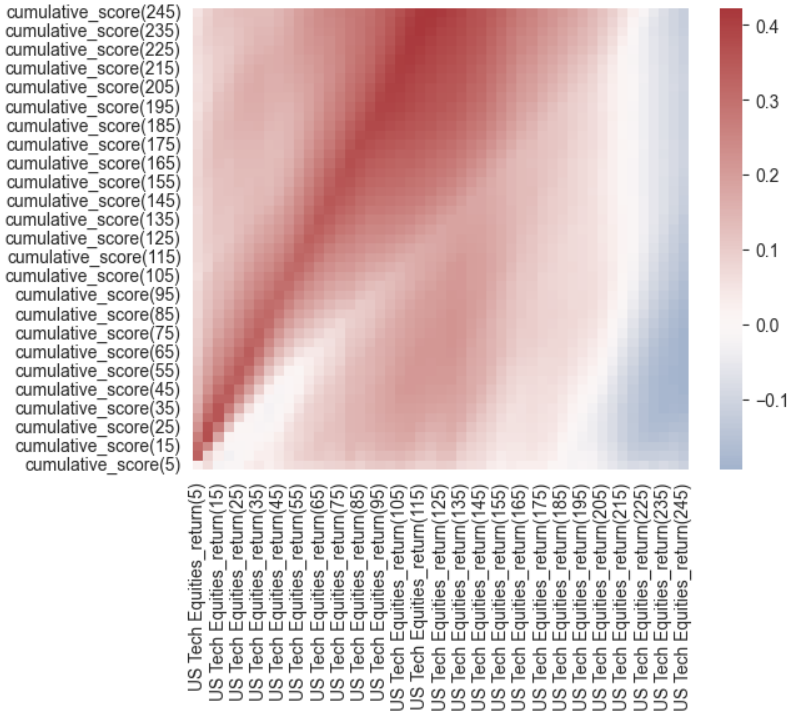

Figure 25: Spearman correlation between USTech and the cumulative sentiment score

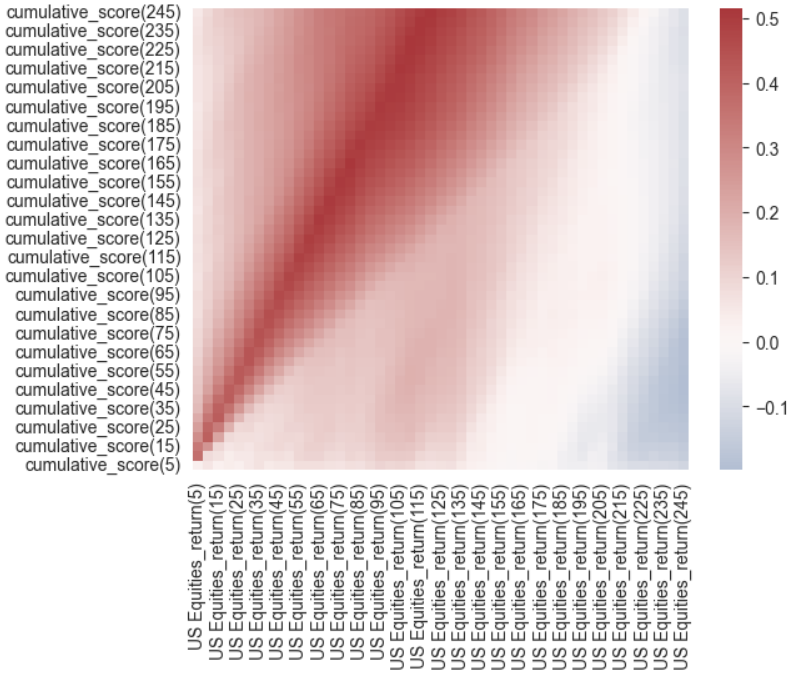

Figure 26: Spearman correlation between the U.S. and the cumulative sentiment score

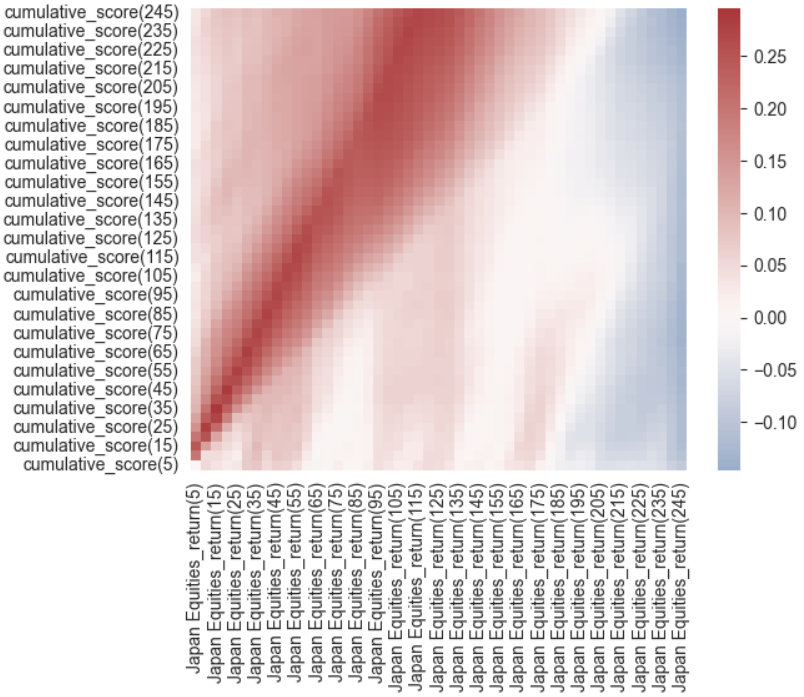

Figure 27: Spearman correlation between Japan and the cumulative sentiment score

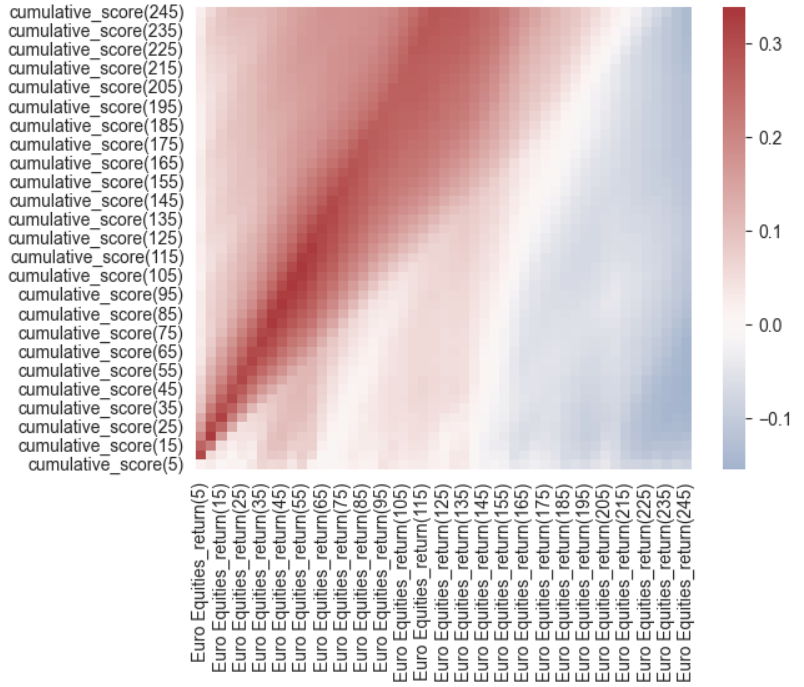

Figure 28: Spearman correlation between Euro and the cumulative sentiment score

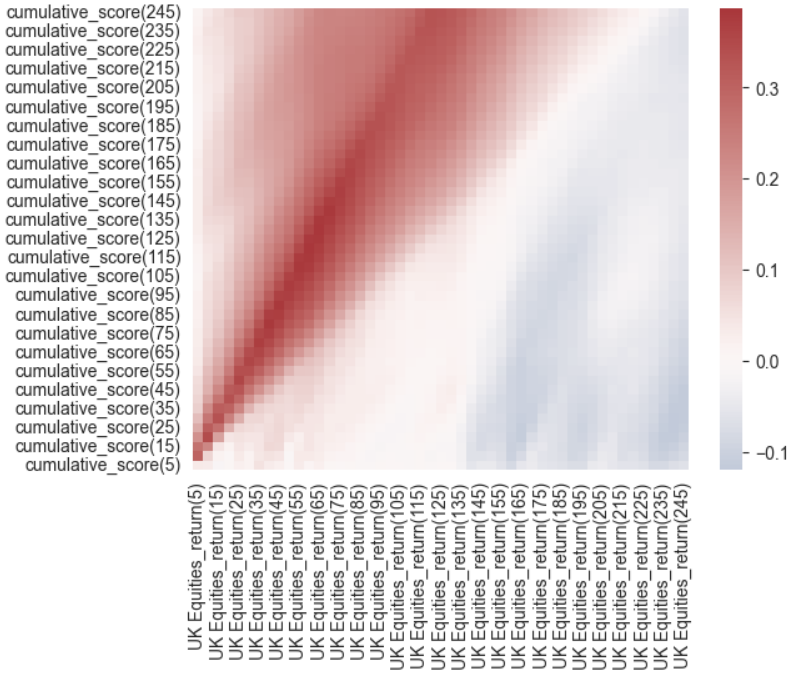

Figure 29: Spearman correlation between the UK and the cumulative sentiment score

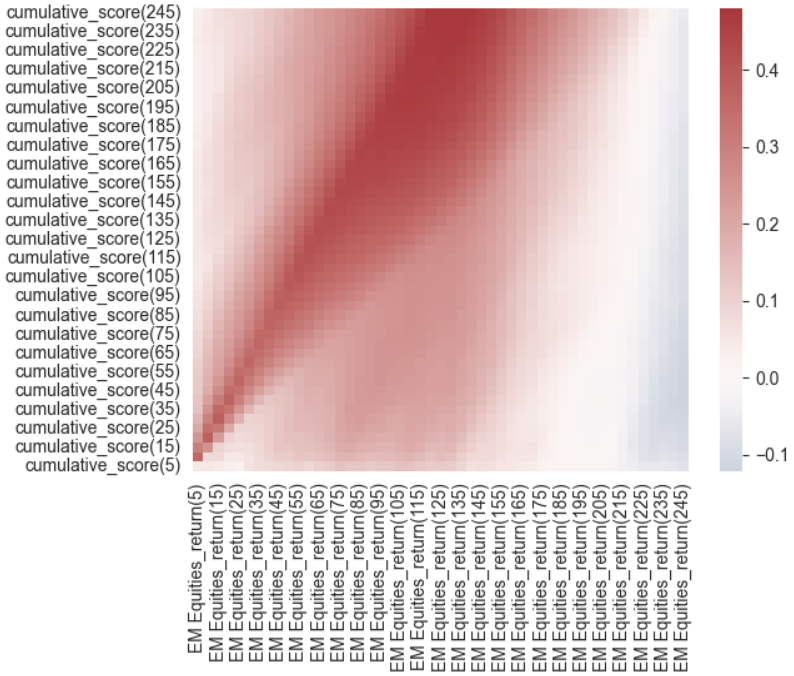

Figure 30: Spearman correlation between EM and the cumulative sentiment score

A.1.5. P-value Spearman Correlation Results

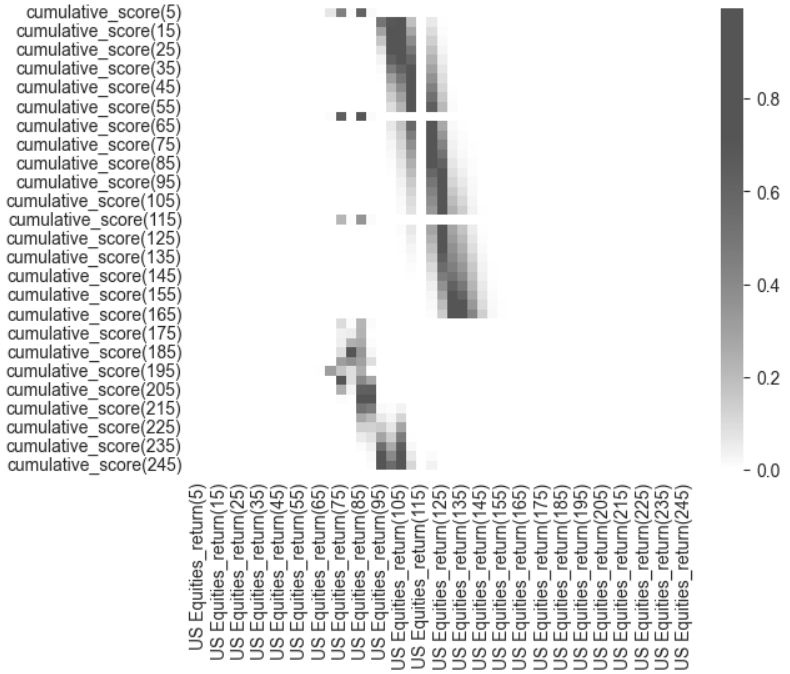

Figure 31: P-value for Spearman correlation between the U.S. and the cumulative sentiment score

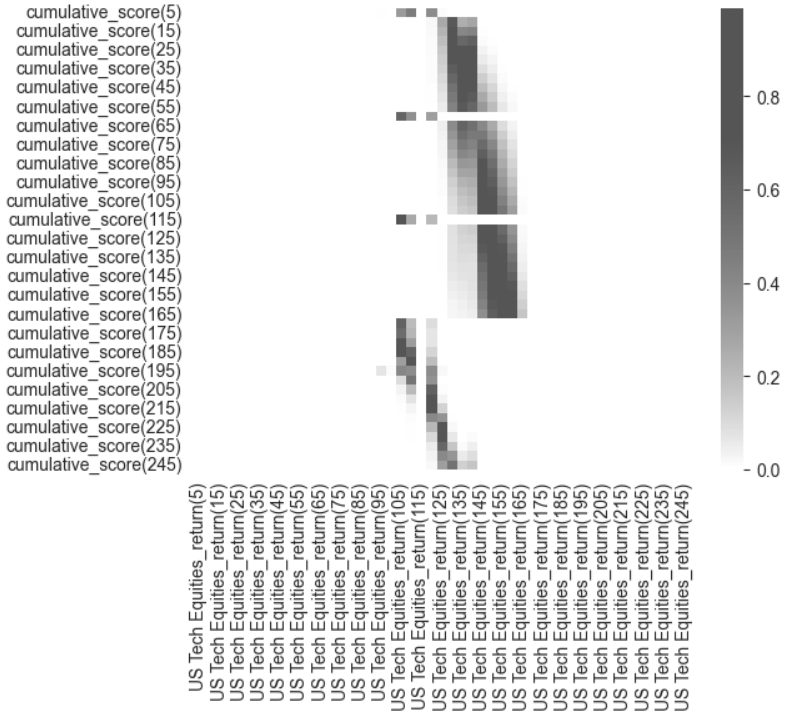

Figure 32: P-value for Spearman correlation between USTech and the cumulative sentiment score

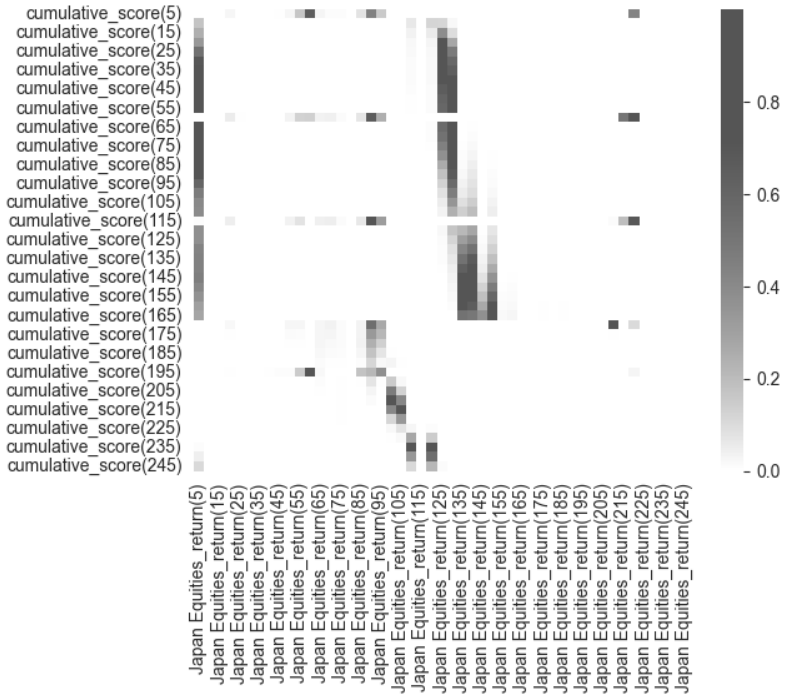

Figure 33: P-value for Spearman correlation between Japan and the cumulative sentiment score

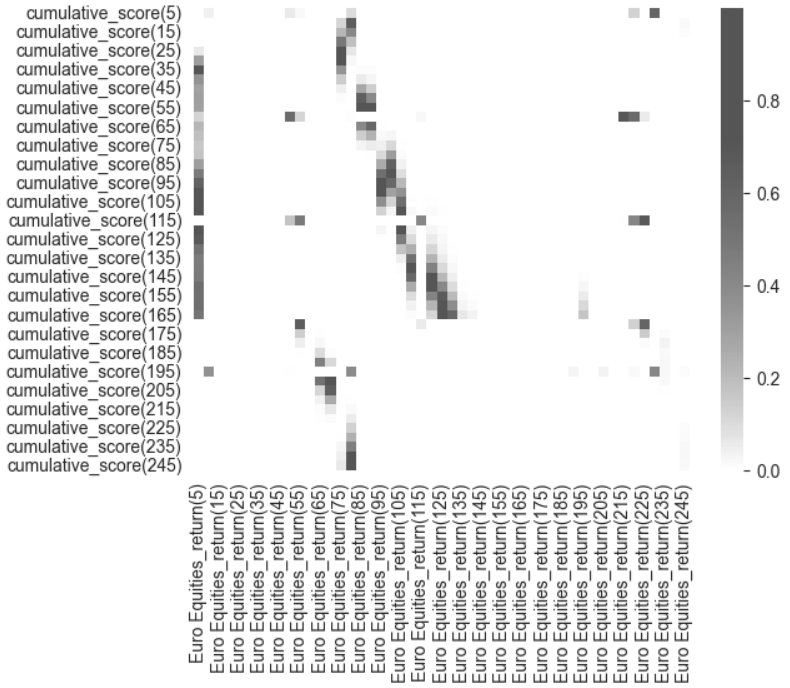

Figure 34: P-value for Spearman correlation between Euro and the cumulative sentiment score

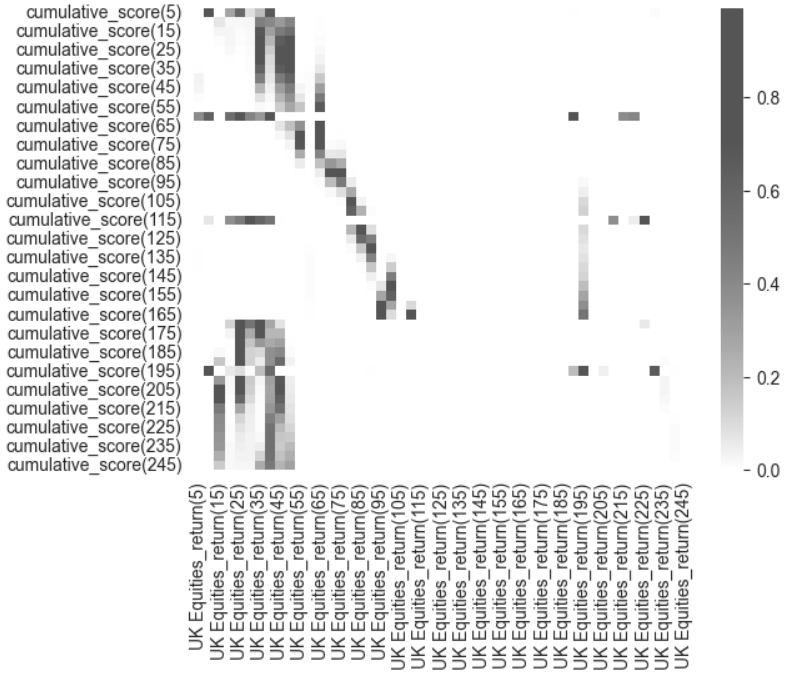

Figure 35: P-value for Spearman correlation between the UK and the cumulative sentiment score

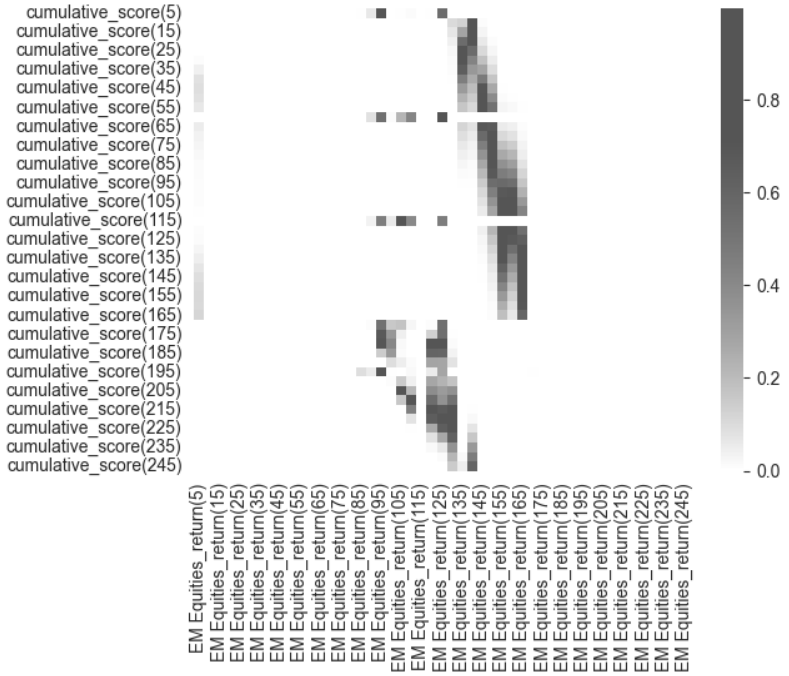

Figure 36: P-value for Spearman correlation between EM and the cumulative sentiment score

A.1.6. Mitigated Spearman Correlation Results

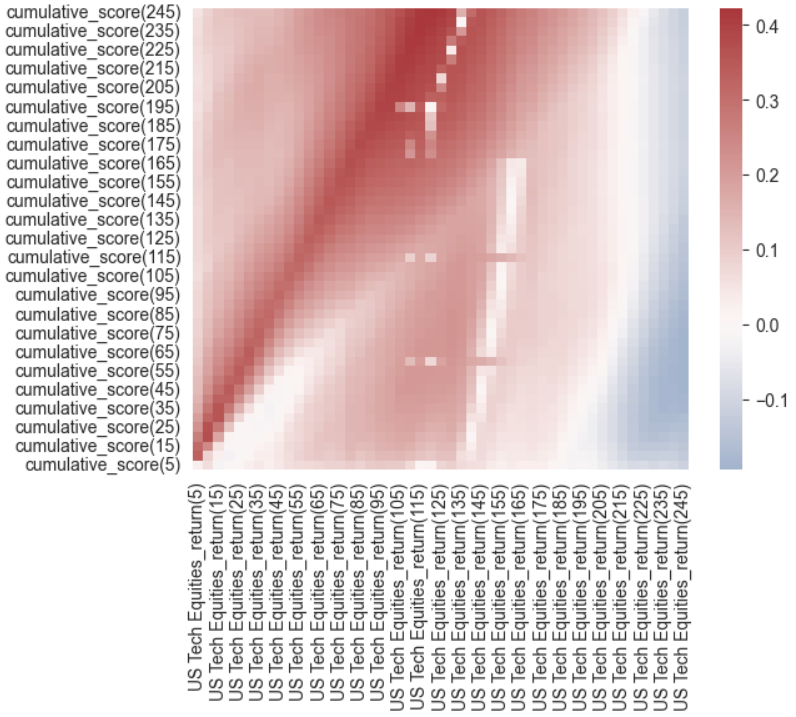

Figure 37: Mitigated Spearman correlation between USTech and the cumulative sentiment score

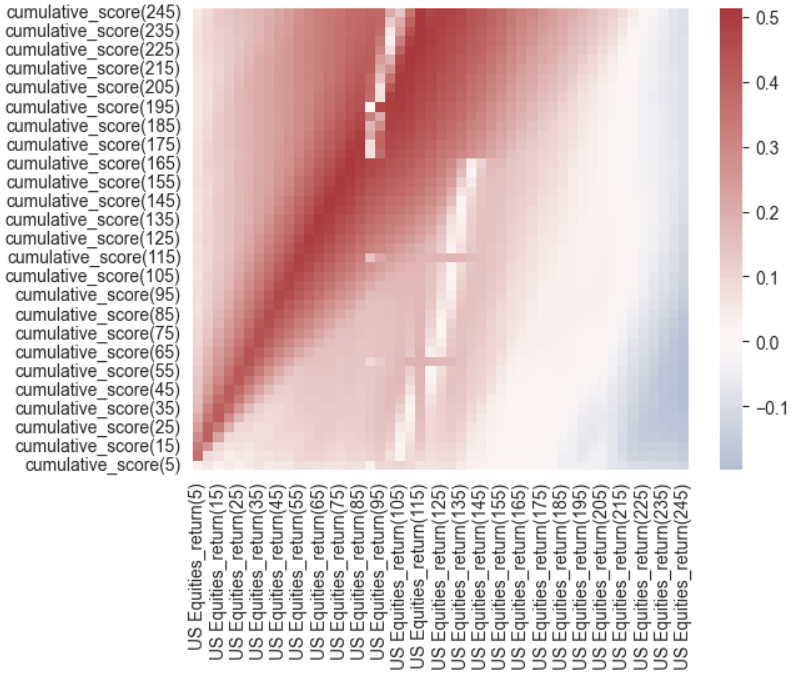

Figure 38: Mitigated Spearman correlation between the U.S. and the cumulative sentiment score

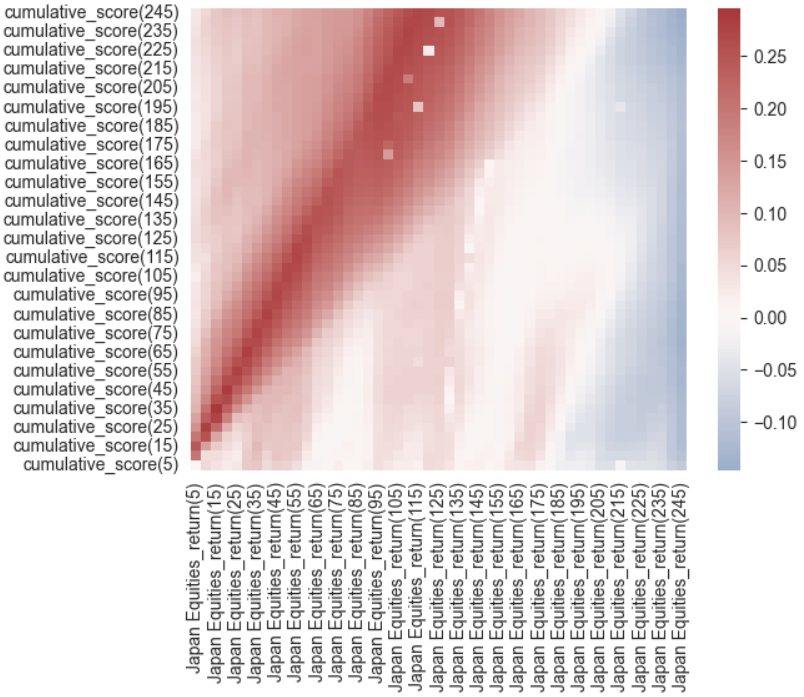

Figure 39: Mitigated Spearman correlation between Japan and the cumulative sentiment score

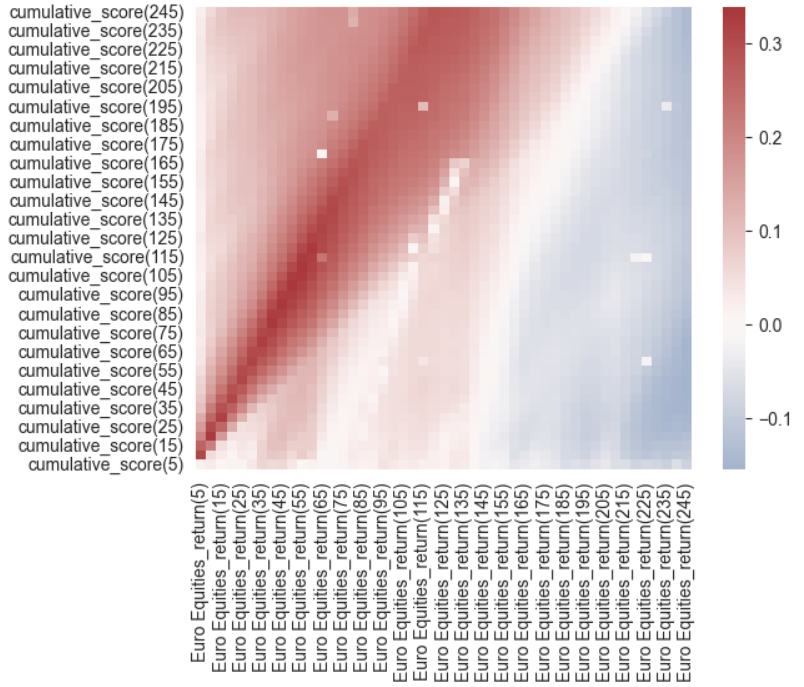

Figure 40: Mitigated Spearman correlation between Euro and the cumulative sentiment score

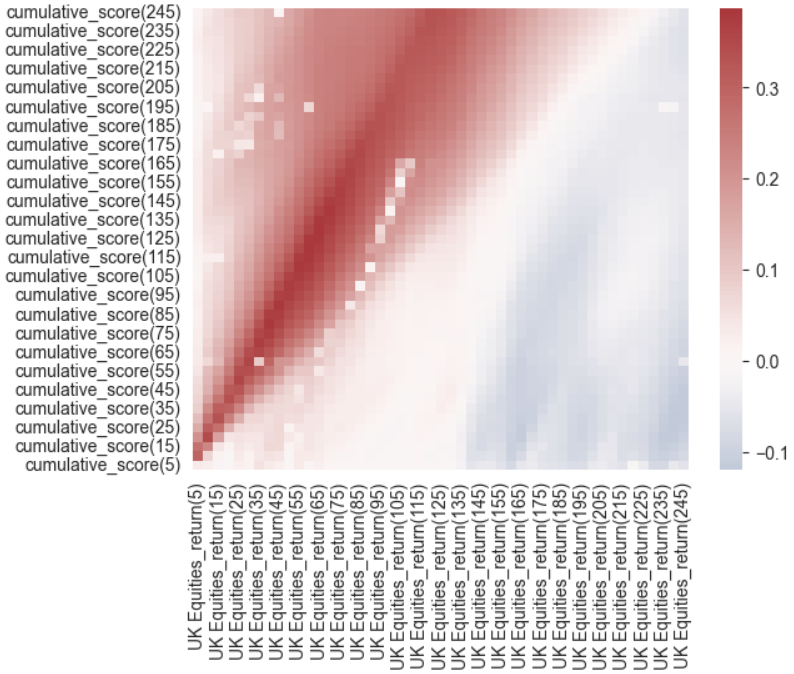

Figure 41: Mitigated Spearman correlation between the UK and the cumulative sentiment score

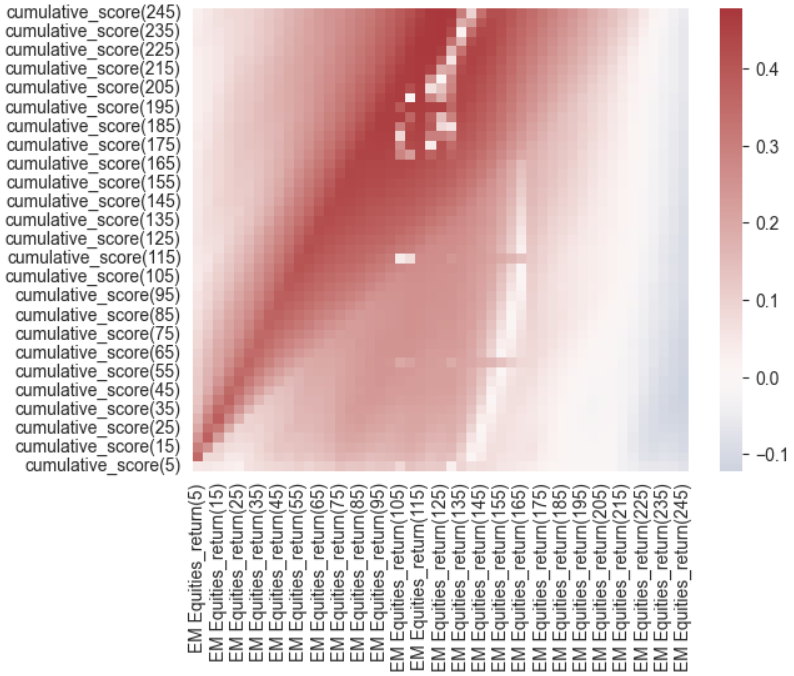

Figure 42: Mitigated Spearman correlation between EM and the cumulative sentiment score

A.2. Optimal Point Determination for the Cumulative Score Lag-Value

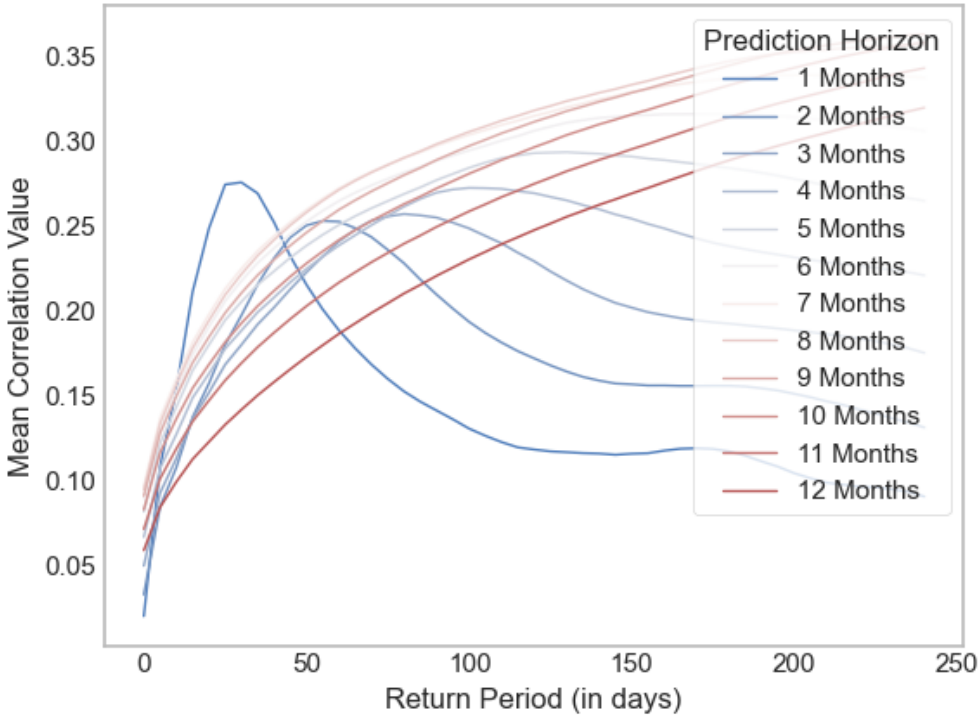

Figure 43: US Tech: Correlation of cumulative score lag over time.

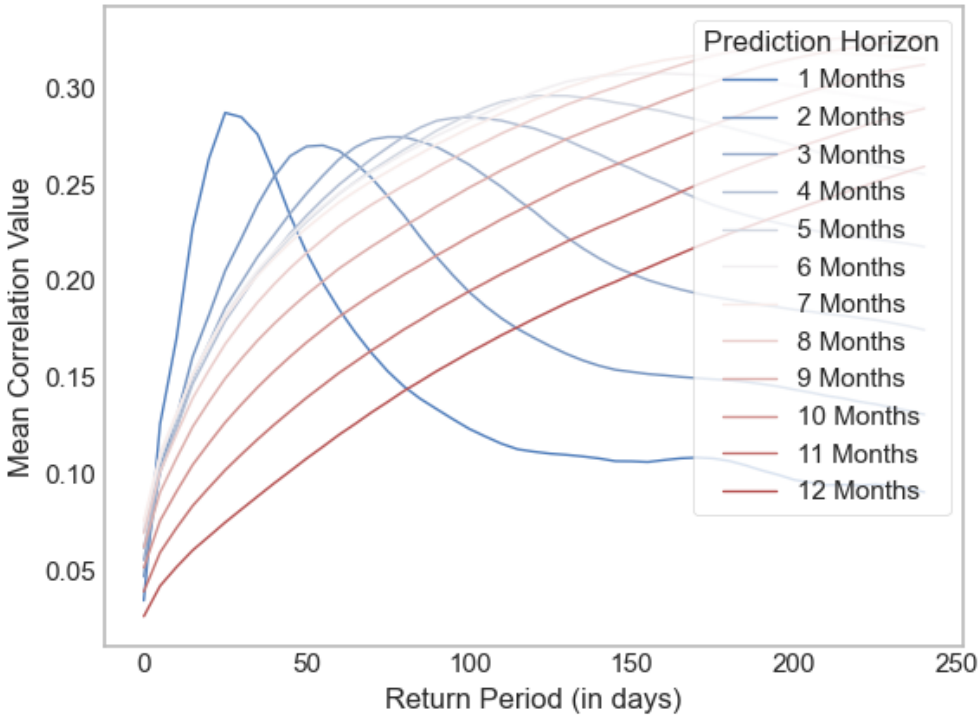

Figure 44: US: Correlation of cumulative score lag over time.

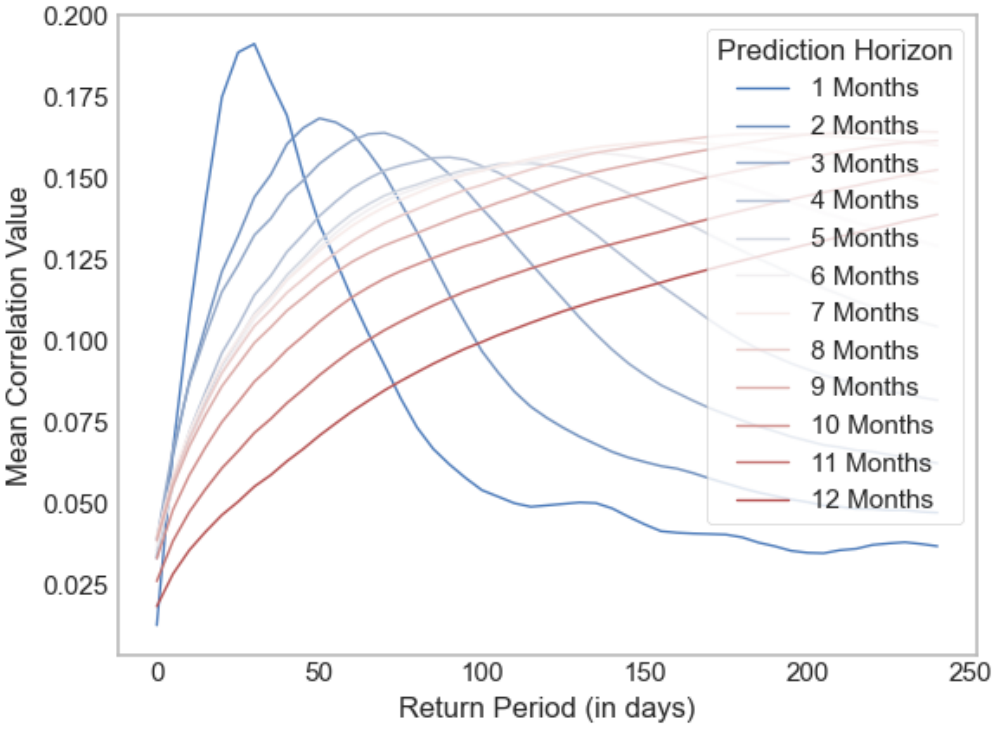

Figure 45: Japan: Correlation of cumulative score lag over time.

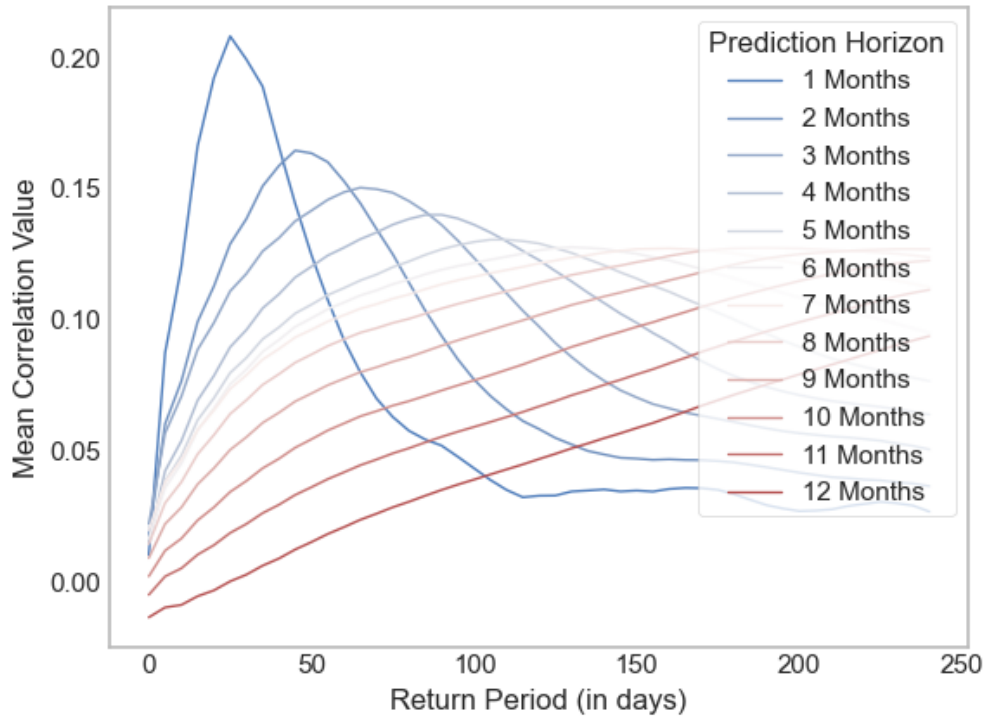

Figure 46: Euro: Correlation of cumulative score lag over time.

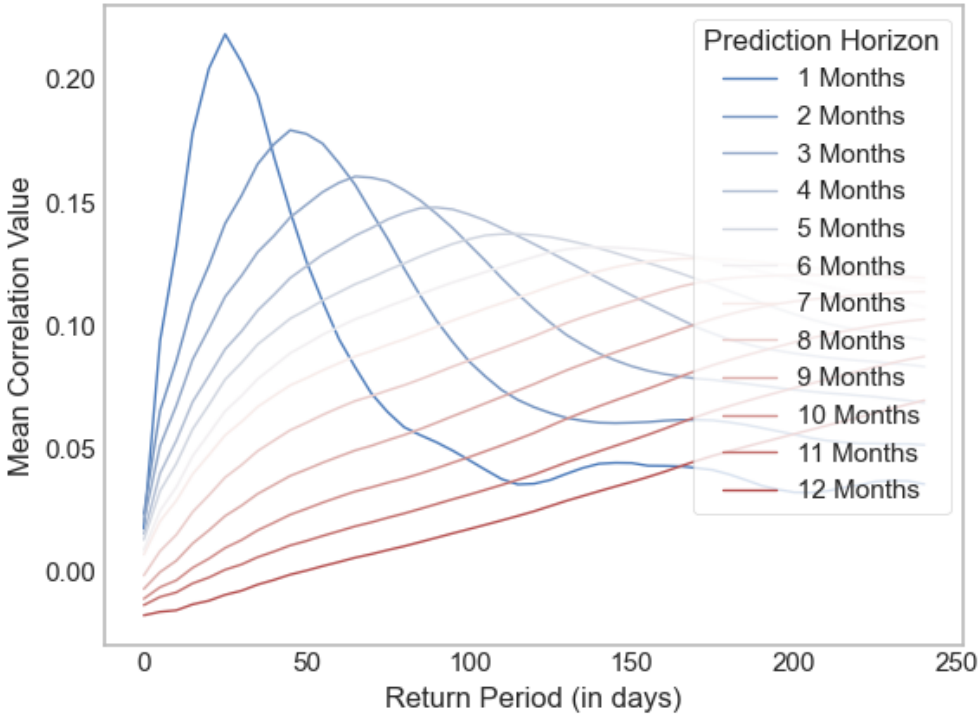

Figure 47: UK: Correlation of cumulative score lag over time.

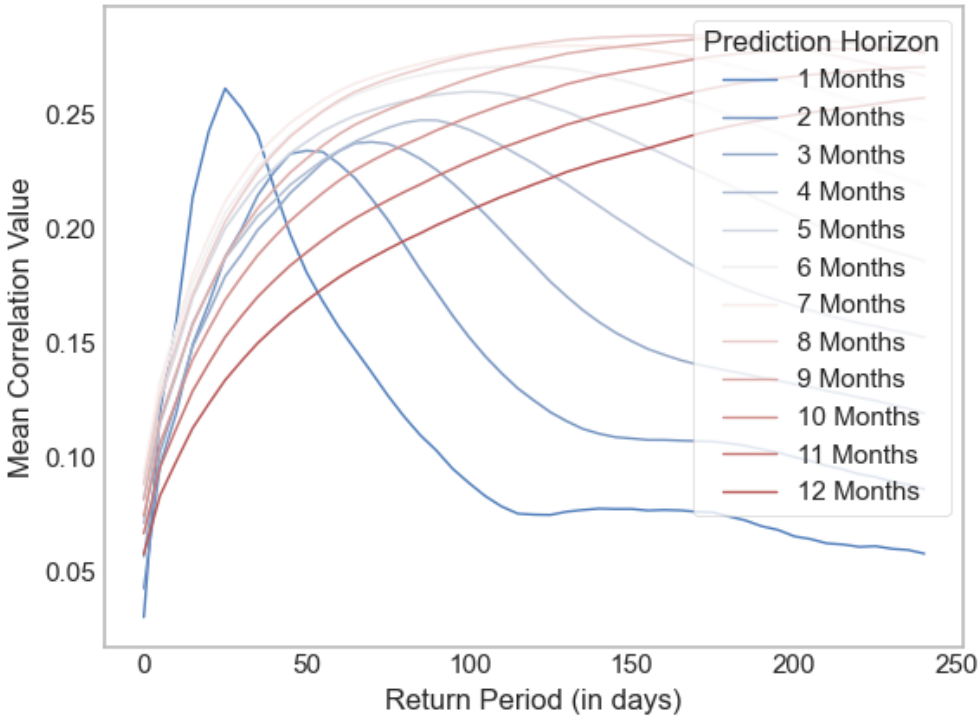

Figure 48: EM: Correlation of cumulative score lag over time.