# Do Elon Musk's Tweets Influence The Market?

Authors: Farouk Boukil, Johnny Borkhoche, Safae Azeroil, Abiola Adeye
Ecole Polytechnique Fédérale de Lausanne
*May* $15^{th}$*, 2023*

***Abstract*—Elon Musk is nowadays one of the most impactful and influential businessmen, so much so that one may suspect that his public declarations may affect the prices and returns of many securities that he publicly criticizes or supports.**
**In this paper, we investigate the predictive power and broadly the impact of his tweets, between 2010 and 2023, on daily asset returns. We determine the presence or absence of such an impact and quantify it when it exists, using two different approaches based on natural language processing and time-series modeling techniques.**

## I. Introduction

Elon Musk has been countless times accused of manipulating the market through his tweets. In 2018, Elon Musk sent forth tweets such as "I can get Tesla share price to 420", effectively increasing the company's stock price. Similarly, Elon Musk's declaration *"ok no more deal"* during Twitters' acquisition attempt in 2022 caused a substantial decline in the latter's stock price.

Our aim is to determine whether there is a causal effect between Elon Musk's tweets and the change in the stock prices of the assets he tweets about, through their daily asset returns. To that end, we use two approaches which converge to similar results. After identifying the publicly traded stocks Elon Musk tweets about by combining algorithmic extractions and named entity recognition, we assess the presence and then significance of the causal relationship based on sentiment analysis and based on an observational study using Bayesian structural time-series modeling.

## II. Data Acquisition and Processing

### A. Data Source and Format

We've conducted an extensive data scrape of approximately 24,000 tweets from Elon Musk, spanning from January 2010 to April 2023, employing Twitter's public API. This rich dataset features over 50 diverse fields, including the text of the tweet, the associated date, the author's ID, a list of users who engaged with the tweet through reactions or retweets, the tweet origin location, incorporated hashtags, and much more.

### B. Data Processing

The original dataset contained numerous fields that were not pertinent to our study, and these were subsequently discarded. We curated a focused dataset with the following integral fields:

1) **ID**: The unique identifier of each tweet.
2) **Created_at**: The timestamp indicative of each tweet's creation.
3) **Text**: The unprocessed text content of each tweet.
4) **Hashtag**: The hashtags embedded within each tweet.
5) **Mentions**: The users identified within each tweet.
6) **Cashtags**: The financial securities delineated by their ticker symbols within each tweet.
7) **Is_reply**: A boolean attribute determining whether the tweet served as a reply to another tweet.
8) **Engagement Metrics**: A range of engagement markers including retweet_count, reply_count, like_count, quote_count, and impression_count, providing insight into each tweet's engagement and interaction levels.

Finally, we transformed the data into a time series table. Each tweet is now indexed by its creation date, making it easier to map each tweet to the time it was posted.

## III. Tweet Exploration

### A. Tweet Statistics

As in most cases, one must reason about the data acquired and manipulate it to extract relevant information suitable for our purpose. This is where data analysis and exploration comes into play.
As explained in the previous section, we've scraped the entire (at the time of writing this) $24'085$ Elon Musk tweets available on his twitter account *@elonmusk*. More specifically, our dataset is made up of $3'673$ original tweets, comprising $15\%$ of the dataset and $20'412$ replies accounting for the remaining $85\%$. This suggests that Elon's main activity on twitter is replying to others. He also seems to be tweeting more over time as seen in Figure 1.

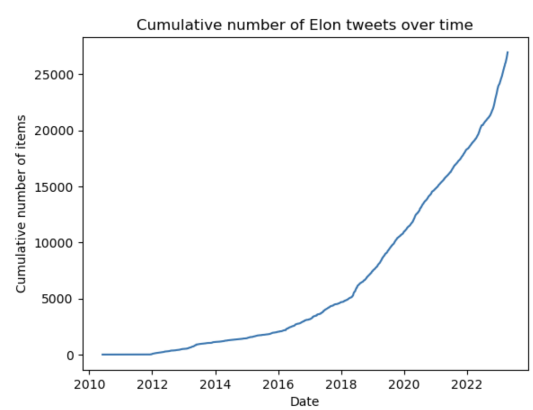


Fig. 1: Cumulative number of tweets over time

We look at some basic statistics regarding the tweets: the mean length of a tweet tallies up to 12.7 words, while the median sits at a comfortable 9 words. Looking at the distribution of tweet length (Figure 2), it is clear that Elon prefers shorter sentence for original tweets but is more articulate when it comes to replies; probably because he usually clarifies things in replies.

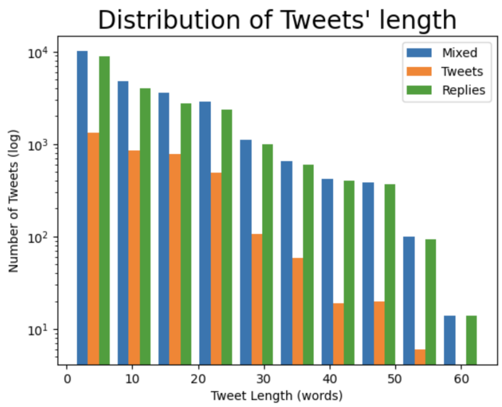


Fig. 2: Distribution of Tweet length

Figure 3 allows us to better visualize the relationship between tweet length and tweet time since his first tweet in 2010. It suggests that the average length of tweets is decreasing over time, despite Elon tweeting more as seen earlier.

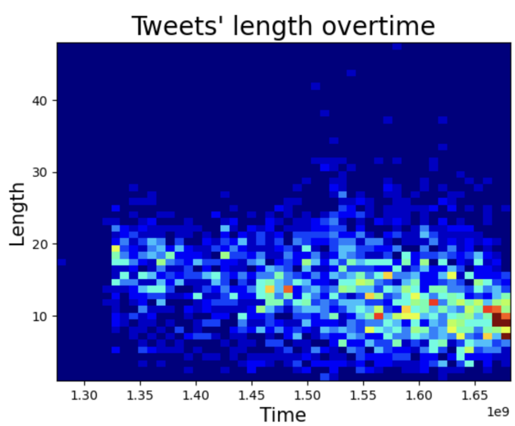


Fig. 3: Distribution of Tweet length over time

By plotting the average daily tweet length over time (Figure 4) we see that there is indeed a negative trend line in the average length of tweets. More specifically, the negative slope trend line is more pronounced for original tweets than for replies, implying that there is no Simpson's paradox at play here. We omit the plots to reduce redundancy.

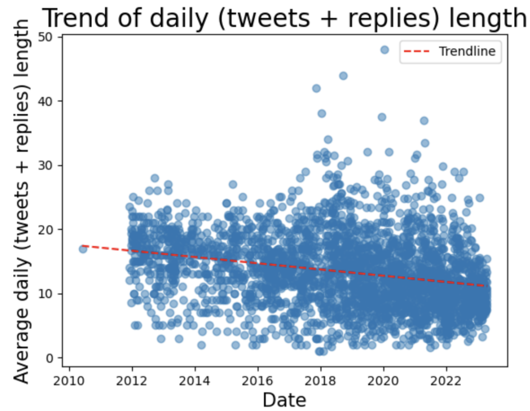


Fig. 4: Distribution of Tweet length over time

Our ultimate goal is to determine whether tweets can be leveraged to predict market returns. Thus, it makes sense to study whether the information disseminated on Twitter has any impact or correlation with the returns. Following this train of thought, highly influential tweets that are seen by more people are more likely to have an effect on the market. We thus observe the mean daily metrics of Elon musk's tweets over time. There are a couple of things to keep in mind; the *impression_count* metric has only become available recently and can be disregarded, tweets before 2014 are very noisy and could thus be disconsidered in our following approaches.

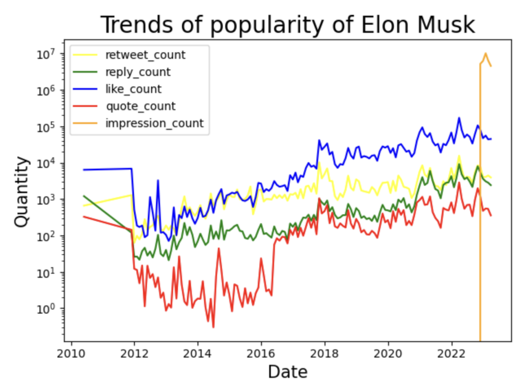


Fig. 5: Distribution of Tweet length over time

### B. *Identifying the Susceptible Market via NER*

So far, we've talked about how we wish to evaluate the predictive power of Elon Musk's tweets on the market - without actually defining said market. Realistically, it would be very difficult to establish or disprove a causal relationship between the tweets and the worldwide markets that span multiple asset classes, securities and countries. Even confining ourselves to a U.S. stock market index such as the S&P 500 remains too broad, thus we will be restricting ourselves to the companies and/or currencies that Elon tweets most often about.

To discover the companies that Elon talks about most often, we first preprocess all of his tweets (original and replies). Text preprocessing is a critical step in natural language processing and machine learning tasks, and it involves transforming raw text data into a structured format that can be analyzed and understood by algorithms. Some common techniques for text preprocessing include tokenization, lower casing, stop words removal, and stemming.

- Tokenization: This is the process of breaking up text into individual words or tokens. It is a crucial step in natural language processing, as it enables algorithms to analyze text at a granular level. Tokenization makes it easier to count words, identify patterns, and perform other analyses on the text.
- Lower casing: This step involves converting all text to lower case. This helps to eliminate inconsistencies in the text data and reduces the dimensionality of the data. This makes it easier to perform text analysis, as algorithms can treat words in lower case as identical regardless of where they appear in the text.
- Stop words removal: Stop words are common words such as "the," "and," "a," and "in" that occur frequently in text but do not carry much meaning. Removing these words from the text can help to reduce the dimensionality of the data and improve the accuracy of text analysis. Stop words removal can also help to speed up the processing of text data.
- Stemming: This technique involves reducing words to their base form or stem. For example, the word "running" might be stemmed to "run." Stemming can help to reduce the dimensionality of the data and improve the accuracy of text analysis by grouping together words that have similar meanings.

Then, we employ a BERT-based Named Entity Recognition model (NER) taken from *HuggingFace [1]* to identify the organizations that appear most often. We kept the most frequent and interesting ones, namely *Twitter, Tesla, SpaceX, Bitcoin, Dogecoin*.

To support out claim, we also looked at the most frequent words that appear in Elon's preprocessed tweets in Figure 6.


Fig. 6: Most frequent words in Elon tweets

Indeed, the top words that appear are mostly related to *Tesla, Twitter, SpaceX* with some of the lower ranked words relating to cryptocurrencies like *Bitcoin* and *Dogecoin*. Unfortunately, *SpaceX* is a privately held company, as such there is no available data on its returns, so we chose to discard it. We thus define the four remaining entities to be the constituants of our market. In the next section, we will formalize the study of tweet predictive power on the market.

## IV. The Influence of Tweet Sentiment on Market Returns

The initial approach to understanding the influence of Elon Musk's tweets on market returns is by analyzing whether a correlation exists between tweet sentiment and subsequent changes in market returns. Specifically, we question whether positive or negative sentiments expressed in Musk's tweets about a specific company or asset translate into a corresponding increase or decrease in the return of that company's stock or asset.

### A. Methodology

To approach this study systematically, we begin by formally defining the problem. We represent a tweet on day $d$ as $T(c, s, d)$, where $c$ refers to the company or asset discussed in the tweet, and $s \in \{+1, 0, -1\}$ signifies the sentiment of the tweet, defined as positive $(+1)$, neutral $(0)$, or negative $(-1)$.

We then establish a window with a pre-effect and post-effect period. This window corresponds to the number of days we analyze before and after the day of the tweet.

In light of this window, we construct two time series of compound returns between days $d$ and $d-$ pre-effect-window and between days $d+1$ and $d+$ post-effect-window for each tweet $T(c, s, d)$. For each company or asset $c$ and each sentiment $s$, we investigate the tweet's impact on returns through hypothesis testing. Our null hypothesis suggests that the returns before and after the tweet follow similar distributions. If the resulting p-value of this test is sufficiently small, it suggests that the tweet likely had a significant impact on market returns.

We then calculate the percentage of tweets with a statistically significant p-value (we use a threshold of 0.05 in our case) for each sentiment and security separately.

### B. Selection of an Appropriate Test

One may consider employing the classical t-test as a hypothesis test to determine if a tweet has had an impact on returns.

The requirements for the t-test are:

1) Data follows a normal distribution.
2) Data points in the pre-effect and post-effect samples exhibit equivalent variance.
3) The absence of autocorrelation in the data.

However, when dealing with stock returns, these assumptions may not necessarily hold. Consequently, we need an alternative test. The Wilcoxon Rank Sum test, a non-parametric hypothesis test, is more suited to our data and requirements.

Unlike the t-test, the Wilcoxon Rank Sum test does not require the data to adhere to any particular distribution, nor does it impose constraints on the variance of the data.

The Wilcoxon Rank Sum test aims to ascertain if the two samples originate from the same distribution by ranking the combined dataset (including both the pre-effect and post-effect samples) and then comparing the sum of the ranks for each sample. If the null hypothesis is accurate (i.e., the two samples derive from identical distributions), we anticipate the sums of the ranks to be roughly equivalent for both samples. Conversely, if the sum of ranks noticeably differs between the two samples, we reject the null hypothesis, suggesting that the tweet has likely influenced market returns.

### C. Results

For our hypothesis testing framework, we employed the following parameters:

- Sample set of securities:
  1) Stocks: TSLA, TWTR
  2) Cryptocurrencies: BTC, DOGE
- Pre-effect window duration: 14 days
- Post-effect window duration: 7 days

The findings indicate a tangible impact of Elon Musk's tweets on the returns of the studied securities:

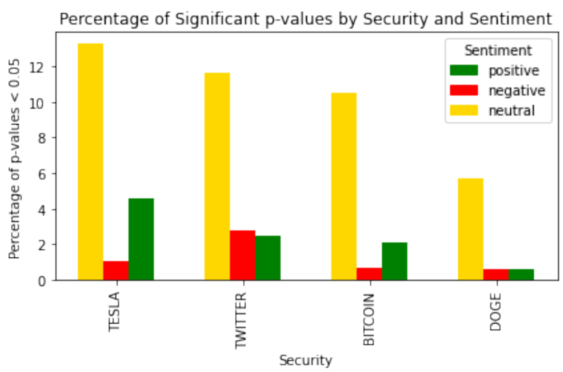


Fig. 7: Impact of Elon Musk's Tweets on Securities

These promising results confirm that Elon Musk's tweets exert a significant influence on the market. However, these findings do not permit us to deduce a specific directional influence on securities' returns.

The primary question we sought to answer was whether a specific sentiment in the tweets correlates with a particular directional move. The answer is no. By examining the tweets that had a tangible impact on the returns, we found no discernable correlation between the tweet's sentiment and the specific directional movements in the security's returns.

We propose several hypotheses as to why this might be the case:

1) A majority of Elon Musk's tweets exhibit neutral sentiments. A tweet with a neutral sentiment could imply both positive or negative news about the security it mentions.
2) Sentiment analysis is not always precise: A tweet with a negative sentiment that mentions a specific security does not necessarily imply negative news about the security (or the company behind it). For instance, Elon Musk could tweet negatively about a Tesla competitor, which could potentially be beneficial for Tesla. This highlights the challenge of automatically capturing these subtle distinctions with our sentiment analysis models.
3) Timing of Tweets and Market Reactions: The markets might react to the sentiment expressed in a tweet at different time scales. For instance, an immediate reaction might be based more on sentiment, while the long-term reaction could be more influenced by the actual informational content of the tweet and other evolving market conditions. This discrepancy could obscure the correlation between tweet sentiment and directional movement in security returns.

In conclusion, while our initial approach confirmed a significant impact of Elon Musk's tweets on market returns, it failed to establish a direct correlation between tweet sentiment and specific directional movements in security returns. This suggests that our method could benefit from further refinement or a different approach, incorporating other factors such as context, content, timing, or more nuanced sentiment analysis techniques.

## V. Bayesian Structural Time-Series (BSTS) Causality Study

Our second approach is rather time-series based. We use a Bayesian Structural Time-Series model, BSTS for short, as presented in the paper *Inferring Causal Impact using Bayesian Structural Time-Series Models*[2] to study the causal effect of Elon Musk's tweets on the returns of the assets of interest.

### A. Theoretical Background & Approach

To fully grasp the use of BSTS for causal inference, it is fundamental to be familiar with the following terminologies:

- The effect or treatment is the event that we suspect to have a causal impact on the observables. In this case, the event is the tweet and the observables are the returns of one or many assets. We do not consider portfolios of these assets, but rather perform our analysis on each one independently.
- The pre-effect or pre-treatment period is a window that is adjacent to the effect's occurence instant, comes right before it and does not include it.
- The post-effect or post-treatment period is a window that is also adjacent to the effect's occurence instant, but comes right after it and includes it.
- The forecast of the observable, generated by an underlying model trained to fit the pre-effect window's observations, is called the counterfactual.

To use BSTS for causal inference, one should construct a synthetic control process assumed to be unaffected by the effect in both the pre-effect and post-effect periods. This synthetic control could be a combination of many real control processes such that it has a high predictive power of the observable, in the absence of the effect, to generate a plausible counterfactual after the effect. In fact, using the synthetic control, one can forecast the "what-should-have-been" (ie. the counterfactual) asset returns in the post-effect period and compare it to the realized return in that same period. Given the high predictive power of the synthetic control, which is measured by its goodness-of-fit on the realized returns in the pre-effect period, a large divergence between the counterfactual and the realized returns in the post-effect period may indicate the presence of a causal impact by the effect on the observable. We go even further and estimate that impact and our uncertainty about it.

This technique requires two assumption, both of which are implied by the choice we made of control processes:

- The control processes must be highly correlated with the observation.
- The correlation link must persist going from the pre-effect to the post-effect period.

To put this into context, our synthetic control is a factor model with 6 factors based on the market risk premium and the following long-short portfolios: HML, SMB, UMD, RMW and CMA. The control processes are the factor risk premia. This factor model explains relatively well the cross-section of expected asset returns and our evaluation of its goodness-of-fit on the pre-effect period for the assets we worked on shows that it is sufficient for our analysis. Although our controls may be influenced by the effect in the post-effect period because they may be portfolios containing our studied asset, the simplest one being the market itself, we neglect that possibility since we consider that the long-short risk premia portfolios and the market portfolios are very diversified so that the weight of the studied asset is sufficiently low.

Bayesian Structural Time-Series modeling relies on a regression with static coefficients used to average the control processes to construct the first component of the synthetic control. In financial terms, we are dynamically building the factor model on the pre-effect period with every different pre-effect period and asset.

In addition to that, the second component is a couple of local long-short term trend and a seasonality state-space models to respectively account for local and long-term trends, as well as seasonality. The justification is that the risk premia are not meant to pick up on these trends, but they may be necessary to construct a good counterfactual if those trends are not identified and mitigated in the data.

It would be intuitive to rely on Maximum Likelihood Estimation (MLE) to fit the training asset returns from the pre-effect period, as one implicitly does when minimizing the mean squared error (MSE) in linear regression. However, this would provide only a pointwise optimal estimate of the parameters we want to learn. The goal of this Bayesian approach is to estimate the uncertainty around these parameters and state vectors, that is, to estimate the joint posterior distribution of the learnable $\beta$ parameters and states $\theta$ underlying the state-space models, conditional on the observed asset returns $r$ in the pre-effect window $w$ of an effect happening at time $t$:

$$P(\theta, \beta | r[t-w:t-1])$$

This is the Maximum A-posteriori Probability rule (MAP). Since this distribution is uknown, the model constructs a Gibbs sampler on a Markov Chain whose stationary distribution is the aforementioned posterior distribution. Gibbs sampling will eventually converge to that stationary distribution at which time sampling the optimal parameters becomes the most likely and we will have obtained a Gibbs sampler from the posterior distribution to estimate it and therefore the uncertainty about our parameters.

Finally, to determine the causality probability, the distribution of the causal effect estimated from the divergence between the actual returns and the counterfacual is statistically tested against a distribution of effects generated under a null hypothesis of no effect.

### *B. Analysis*

Similarly to the previous analysis, we focus on the most popular tweets by number of likes and by sentiment provided by BERTweet. In fact, we believe that unpopular tweets have significantly smaller odds of having a real impact on the market. On the other hand, popular tweets gain more visibility both on and off Twitter thanks to shares and retweets but also thanks to Twitter's feed algorithm that would tend to show trending tweets more often, which helps quickly propagate the information to the market actors like investors, shareholders and institutions.

Let us consider the following positive sentiment tweet, which is in this case a positive indicator of the company's attractiveness and growing customer base:

*"Lots of Tesla cars to deliver before year end! Your support in taking delivery is much appreciated."*

Figure [8] shows the actual Tesla daily returns during both the pre-effect window of 14 days and the post-effect window of 3 days, against the forecast in the pre-effect window and the counterfactual in the post-effect window. The figure also shows the pointwise difference in the daily returns between the predictions and the actual values over the entire period. The grey area corresponds to 95% confidence interval of the forecast and the counterfactual.

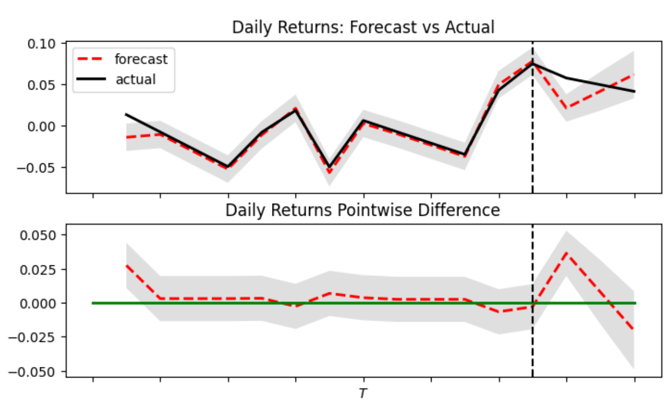


Fig. 8: Impact of a positive tweet on Tesla's daily return

As explained previously, the reason why we take a small window of a few days no more is that on one hand we need to account for the time needed for information to propagate and be incorporated into asset prices and returns, and on the other hand we need to avoid as much as possible the influence of other hidden covariates that may occur in the post-effect window.

It is obvious that the bayesian model has a very nice goodness-of-fit on the returns of the pre-effect window, which further consolidates that the counterfactual should be a relatively good approximation of what should have happened in the absence of the effect as we exploit the predictibility of the daily returns. The causal effect, estimated to be present with no uncertainty in this case, is reflected in the fact that, while the model predicted a sharp overnight drop of the return, the announcement may have amortized the shock during the couple of days that followed this announcement. The point-wise difference plot quantifies the effect by showing the difference between the actual daily returns and the counterfactual ones. Indeed, the actual returns remain above our expectation for around two and a half days.

Although this particular announcement may have, according to our study, had an impact on Tesla's daily return, this impact is not consistent on the cross-section of Tesla-related tweets, nor is it consistent across those sharing the same sentiment. To illustrate, Figure [9] shows the results for a neutral sentiment tweet with causality likelihood of less than $2\%$:

*"Tesla is going to develop a quiet, electric leafblower."*

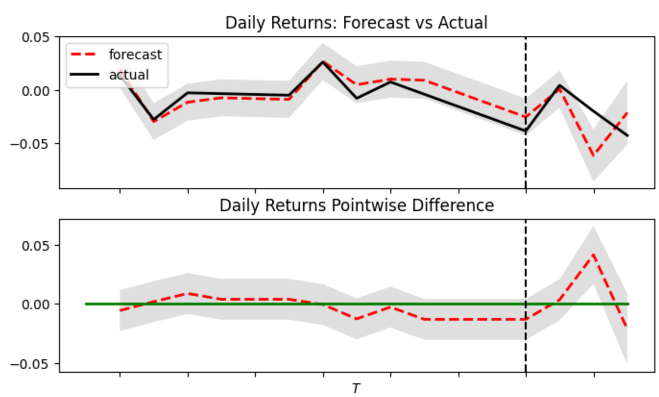


Fig. 9: Impact of a neutral tweet on Tesla's daily return

Even if there is a significant divergence starting day 3, prior to that the counterfactual is very close to the true returns and so the likelihood of a causal impact is very low. In fact, it may be the case that some other hidden co-variate causes the divergence later than the date at which the tweet is posted.

The inconsistency in the causal influence of Elon Musk's tweets is consistent across the securities we studied. For instance, Figures [10] and [11] showcase the analysis of the following tweets about Twitter Inc. and Bitcoin:

*"Twitter s#cks"*

*"[...] I still own and won't sell my Bitcoin, Ethereum or Doge for what it's worth."*

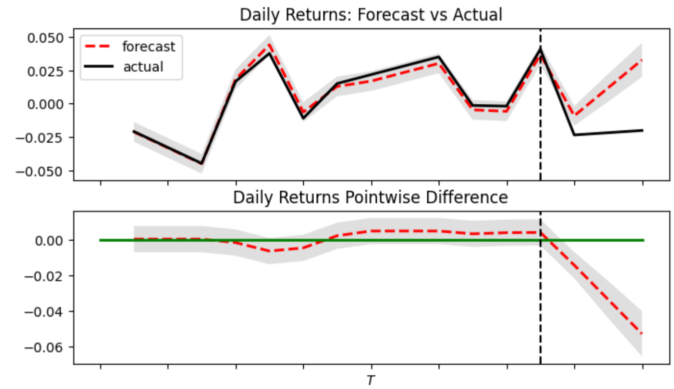


Fig. 10: Impact of a tweet on Twitter Inc.'s daily return

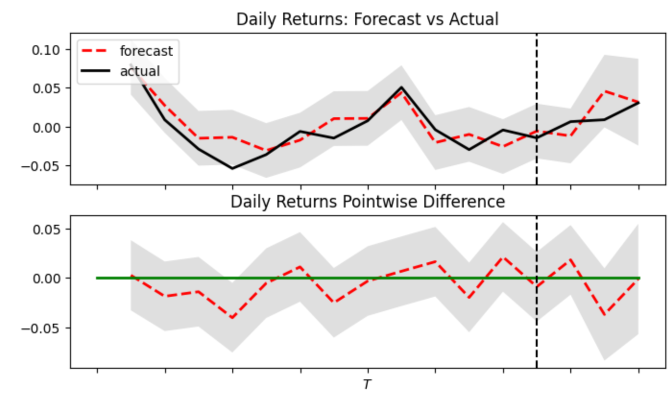


Fig. 11: Absence Impact of a tweet on Bitcoin's daily return

Figure [10] shows that Elon Musk's critic of the social media Twitter may have amplified the negative shock to the daily returns of Twitter Incorporation. Furthermore, they remain consistently below the counterfactual and outside the confidence interval. This persistence, on top of the fact that the model fits well the pre-effect window's daily returns, further strengthens the hypothesis of causal impact. Indeed, the probability of a causal impact is $98.8\%$.

Conversly, Figure [11] illustrates a situation where causal impact is likely to be absent as our uncertainty about the learnable parameters ($\beta$, $\theta$) may explain the divergence between the actual returns and the counterfactual.

Bayesian structural time-series analysis is able to give us insight about the possible existence of causality effects to partially explain the shocks or improvements to daily returns. Nevetheless, it is far from perfect and can suffer tremendously from the hidden covariates problem like in any observational study. All in all, some tweets, independently of their sentiment and of the asset in question, may be associated to a causal impact whose nature does not seem to correlate with the predicted sentiment. We deduce that a few of Elon Musk's tweets may impact the market, though it is hard to isolate his single-handed impact.

## VI. Conclusion

From the sentiment analysis study, we have shown that the nature of the impact, whether it is positive or negative, is hardly predictable. In fact, it does not necessarily correlate with the sentiment of his tweets, nor is this impact always present when we expect it the most. The Bayesian Structural Time-Series Modeling approach has similar conclusions. All in all, Elon Musk can have an impact on the assets he tweets about, but this impact is unpredictible and inconsistent.

## References


[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[2] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott, "Inferring causal impact using bayesian structural time-series models," *The Annals of Applied Statistics*, 2015.