Money Talks, Sentiment Walks: Navigating the News-Finance Nexus

Marcus Vinicius Santos[†]
Fernando Morgado Dias[‡]
Thiago Christiano Silva[*]
Fábio Mendonça[**]

Abstract This study performs sentiment analysis on financial news to assess the impact of sentiments on financial indices using an integrated framework encompassing sentiment analysis and a causal discovery algorithm based on Gaussian process regression to deal with non-linear relationships. Specifically, a news sentiment database was created by extracting financial news articles on various topics sourced from the investing.com web page. The news articles were subsequently inputted into the FinBERT model in order to extract their respective sentiments. The news dataset was combined with financial indicators, specifically the S&P 500 index, dollar index (DXY), and Brent crude oil. Subsequently, the causal discovery model, namely LPCMCI, was employed to conduct causal discovery on the variables of interest. Different from previous works, this study gives a wide perspective on the connections between news and financial behavior, accounting for lag dependencies and the direction of the impacts. The main results indicate a reciprocal relationship between news sentiments and financial variables. In this context, the S&P 500 index can be regarded as a source of Stock Market news together with Commodities news sentiments, where a contemporaneous and lagged effect is observed towards Brent returns. In the opposite direction, news sentiments regarding the Currencies subject appear as a driver of DXY returns. Furthermore, it was observed that various subjects exhibit distinct patterns of interaction with individual financial indicators. The findings of the study clarify what was previously suggested in the literature and provide fresh perspectives that have not been investigated thus far.

Keywords: Finance; Sentiment Analysis; Causal Discovery; LPCMCI; Gaussian Process Regression.
JEL Code: G1, G14, G41, C14, C82.

1. Introduction

In the modern interconnected global economy, financial markets play a pivotal role in shaping the economic landscape. Investors, policymakers, and

Submitted on April 7, 2024.
[†]Catholic University of Brasília (UCB), Brazil: marcus.ssz@gmail.com
[‡]University of Madeira and Interactive Technologies Institute (ITI/LARSyS), Brazil: morgado@staff.uma.pt
[*]Catholic University of Brasília (UCB), Brazil: thiagochris@gmail.com
[**]University of Madeira and Interactive Technologies Institute (ITI/LARSyS), Brazil: fabiorubenn@gmail.com

analysts closely monitor financial indicators to gauge the health and performance of these markets. Simultaneously, the rise of digital media and real-time information dissemination has led to a significant paradigm shift in how financial news is generated, disseminated, and consumed (Sohangir et al., 2018). The literature has been interested in uncovering the role that media news has played on the financial market for a long time (e.g., Oberlechner and Hocking, 2004; Alanyali et al., 2013; Balcilar et al., 2017). One possible reason is that it incorporates the Efficient Market Hypothesis (EMH) formulated by Fama (1970), which asserts that financial markets are highly efficient in processing and reflecting all available information. It could be posited that EMH assumes that investors have rational expectations, meaning they process information correctly and make optimal decisions based on that information. This view suggests that financial news leads to efficient price adjustments.

Several studies have shown a positive relationship between news sentiments and market performance. For instance, Mohan et al. (2019) suggests a correlation between the textual information and stock price direction. Moreover, Alanyali et al. (2013) hypothesizes that movements in the news and movements in the markets may exert a mutual influence upon each other. This phenomenon introduces the conceptual framework of "exogenous and endogenous news," where endogenous news emanates as a direct consequence of the indicators' performance, while exogenous news materializes in response to external factors beyond the scope of the indicators themselves. In an attempt to overcome this issue, Walker (2016) uses two distinct markets: one to source the media coverage and another to test its effect. Since the latter does not influence the former, the impact is "cleaned" from endogenous news.

Besides the intricate relationship between media news and financial indicators, there is also the question of how different news topics interact with the different indicators. Hisano et al. (2013) identified 715 different news topics and their relationship with the trading activity of the stocks of 206 major firms included in the S&P 500 index. Their results showed that large volumes of trading can often be explained by the flow of news. Similarly, Feuerriegel et al. (2016) posits that not all news has the same effect on prices and uses Latent Dirichlet Allocation to determine different topics and their impacts on abnormal stock returns. They found that the topics indeed have different impacts, whereas 'drug testing' exhibited a large effect.

With this in mind, this paper aims to systematically investigate the intricate relationship between different financial news topics and relevant financial indicators, specifically focusing on the S&P 500 index, the dollar index (DXY), and the Brent crude oil as they stand as prominent benchmarks that reflect the collective sentiment and expectations of investors in a wide range of economic sectors. The financial news topics were extracted from the Investing.com web portal under the following issues: Commodities, Stock Market, Economy, Economic Indicators, and Currencies.

To extract the news, a web crawler was designed, resulting in 249,155 daily news stories across the different topics and for the period comprising from June 9, 2015, to March 6, 2023. The composition of news articles consists of a headline and a textual corpus. Then, they were submitted to a sentiment analysis model, namely FinBERT, which returned three possible classifications: positive, negative, or neutral. To mitigate possible erroneous classifications, a summary of the textual corpus was generated with a summarizing language model, and the sentiment model was also applied to it. Then, each news item was reclassified according to the sentiment found more times across the three inputs: headlines, full text, and summary. After constructing the sentiment dataset, it was aggregated to accommodate the daily structure of the financial variables.

In situations where controlled experiments are impractical, the current framework of causal discovery offers an alternative way to learn about cause and effect through observational, i.e., non-experimental data (Gerhardus and Runge, 2020). In this sense, a Python module for time series analysis, namely Tigramite, was used to reconstruct causal graphical models based on the Latent-PCMCI[1] framework. The causal meaning is attributed to the physical mechanisms by which the value of one given variable is determined from the values of a set of given variables, including the one in question.[2] Besides the possibility of establishing a causal relationship between the series, the framework allows for the presence of confounding factors in the model and deals with non-linearity issues, common when dealing with financial returns (Chen and Hao, 2018). Furthermore, the model utilizes conditional independence tests that were performed by Gaussian process regressions, a probabilistic supervised machine learning framework that is useful when dealing with problems where it is required to assess unknown functions that map inputs to outputs and when the shape of the underlying function is unknown (Schulz et al., 2018).

The results have shown a complex relationship between the analyzed vari-

[1] In the work of Jiang and Shimizu (2023), it is explained that the LPCMCI model has its roots on the PC algorithm, named after Peter Spirtes and Clark Glymour (Spirtes et al., 2000), which in turn has been subjected to many variations, such as the Fast Causal Inference (FCI) (Glymour et al., 2019), which is the fundamental basis for the LPCMCI algorithm (Gerhardus and Runge, 2020), where "MCI" stands for Momentary Conditional Independence.

[2] It is important to note that the causal meaning is primarily statistically based and does not encompass an economic causal interpretation.

ables. On the one hand, each financial variable relates differently to each news topic sentiment. News that covers the Stock Market and their respective sentiment are liked primarily by the S&P 500 index. It was observed that contemporaneous and two-day lag volatility of the S&P 500 index leads to changes in the Stock Market news sentiments. A similar behavior is seen between Brent crude oil and Commodities news sentiments, although with a smaller confidence level, and the same is seen between Brent and Stock Market news sentiments. On the other hand, news sentiments under the Currencies topic exhibited a direct link with DXY. Moreover, secondary effects were also found, where the relationships between sentiment variables and financial indices are due to unobserved factors. The results were corroborated by robustness tests such as correlation and Granger causality tests. Furthermore, a different dataset was used to increase the accuracy of the sentiment analysis model.

As elucidated in the existing literature, the results support a converse directionality, wherein financial indicators exercise a discernible influence over financial news. This phenomenon introduces the conceptual framework of "exogenous and endogenous news," where endogenous news emanates as a direct consequence of the indicators' performance, while exogenous news materializes in response to external factors beyond the scope of the indicators themselves. This interplay between financial indicators and news dynamics not only deepens the understanding of the information flows in financial markets but also underscores the complex relationship between these two fundamental elements of market behavior.

This work provides significant contributions to the fields of financial and economic research, as well as to the existing literature in computer sciences, given the close relationship between the methods employed, such as natural language processing, large language models, and machine learning techniques, and the computer sciences field. First, it sheds light on the complex relationship between sentiments from financial news and relevant financial indicators, encompassing different fronts of the US economy, such as the stock market, commodities market, and foreign exchange market. Second, this research leverages advanced methodologies, including Natural Language Processing (NLP), Large Language Models (LLM), and sentiment analysis coupled with the latent PCMCI model and Gaussian process regression. These methodological contributions expand the toolkit available to researchers and set a precedent for future investigations in this area. Lastly, this work also compares the results with methods well established in the literature, such as correlation and Granger causality tests.

Moreover, there is a large stream of literature that is focused on the use of sentiment analysis to forecast financial indicators (e.g., Kordonis et al., 2016; Kalyani et al., 2016; Ko and Chang, 2021). From a different point of view, this study is mainly interested in a wide perspective, where one can evaluate how different news topics, specifically from a world-renowned news data portal, interact with different financial indicators, considering lag dependencies and the direction of the impacts. Additionally, this study utilizes causal discovery models that rely on Gaussian process regression to detect these links. To the best of the authors' knowledge, this is the first study to use these methods on sentiment scores and financial variables. Ultimately, the findings of the study clarify what was previously suggested in the literature and also provide fresh perspectives that have not been investigated previously.

The structure of this paper is organized as follows: Section 2 provides a comprehensive review of the existing literature on the subject, highlighting key findings and gaps in current knowledge. In Section 3, the data and methodology adopted for the analysis are outlined, encompassing the sentiment analysis framework and the causal discovery model. Section 4 presents the empirical findings, showcasing the quantitative insights garnered from the relationship between financial news and the identified indicators. A discussion of these findings in the broader context of financial markets and their implications is exposed in Section 5. Finally, Section 6 concludes the paper by summarizing the main findings, contributions, and potential avenues for future research.

2. Related work

Over the past decades, the financial sector has witnessed a remarkable transformation regarding information availability and accessibility. Traditional sources of financial news, such as newspapers and specialized magazines, have been largely supplemented, and in some cases, supplanted, by digital platforms and social media channels. This rapid evolution has democratized access to financial information, actively enabling individuals and institutions of various scales to engage with market-related news and insights. Consequently, financial news has become not only a source of information but also a driving force behind market sentiment, investor behavior, and ultimately, the trajectories of financial indicators (Blasco et al., 2005; Mishev et al., 2020; Wankhade et al., 2022). Furthermore, Boubaker et al. (2021) specifies that credible financial media is essential when it comes to informing investors. Investor sentiment can be greatly impacted by the news, and this can be seen in the decisions that they make. In the event of sharp volatility in the financial markets, the financial media is supposed to alert the public swiftly, to monitor the situation as it develops, and to produce expert, frequent reporting.

From a different perspective, Walker (2016) claims that determining the extent of media influence on financial markets is challenging. Since there is now 24-hour news, market performance virtually immediately determines how much media coverage there is. Moreover, future market performances may even impact media coverage when moves are highly anticipated. As such, determining media influence cannot be based solely on a correlation between market performance and media coverage. In an attempt to tackle this issue, the author shows that media influence extends beyond the market on which it is reporting and, in doing so, circumvents the causality issues usually prevalent when studying media influence in markets. The author regressed the return premium and trading volume of a portfolio of equities focused on the UK residential property market against the number and optimism of Financial Times housing market articles to test the media's effect, identifying a positive link between the optimism of media reporting and portfolio return premiums.

Boubaker et al. (2021) also emphasize that there is a natural connection between the news and the financial markets, and that they can influence each other. Their research explored the idea of news diversity as a measure of news importance–if numerous outlets report similar events, topics covered by the news will be concentrated. This implies that these occurrences have substantial influence and can significantly impact movements within the financial market. Thus, news diversity closely tracks financial market changes, whereas a more concentrated news diversity index will better reflect market movements. In this context, using a change-point analysis, they examined the relationship between news diversity and the movements of financial markets and assessed the prediction effect of news diversity on market prices and volatility of the financial market. The results revealed that news diversity can be used to forecast the movements of the stock market accurately.

Financial news are also utilized to improve the prediction of macroeconomic indicators, as Feuerriegel and Gordon (2019) proposed by an alternative method in which terms from distinct semantic categories are mapped onto latent structures. Financial news reflects both the present health and future expectations of corporations, according to their findings. For product-related indicators, such as the actual industrial product rate and confidence indices, news-driven models therefore appear to be particularly advantageous. When it comes to long-term forecasts, news-driven models exhibit a tendency to outperform benchmark models. Regarding the commodities sector, Li et al. (2022) utilizes a web crawler to construct a Chinese investor sentiment index to explore the dynamic relationships between shocks in the crude oil price and investor sentiment. Then, a structural vector autoregression model divides crude oil price shocks into three categories, and a wavelet coherence analysis is used to analyze the temporal and frequency domains' dynamic correlation between crude oil price (shocks) and investor mood, as well as their asymmetric dynamic connection under distinct crude oil price trends. The results indicate heterogeneous dynamic correlations and lead-lag interactions between crude oil price (shocks) and investor mood across time and frequency domains. In addition, asymmetric dynamic correlations and lead–lag interactions exist between crude oil prices (shocks) and investor mood under different price patterns. Gong et al. (2022) captured the trend of oil futures prices based on the text-based news where the text features obtained from online oil news catch more hidden information, improving the forecasting accuracy of oil futures prices. Moreover, they found that the textual features are complementary in improving forecasting performance and verifying the asymmetric impact of positive and negative emotional shocks on oil futures prices.

Apart from only establishing a link between news and financial indicators, the interest also lies in the impacts of different topics on the financial series. In this respect, Shynkevich et al. (2016) utilize different categories of news articles simultaneously with a multiple kernel learning technique. Specifically, news articles were allocated to five categories based on their relevance to a target stock and its sub-industry, industry, group industry, and sector. Multiple kernel learning (MKL) techniques were used to learn from these news subsets so that independent kernels were utilized for each subset. Using stocks from the healthcare sector, they found that the highest forecasting accuracy and trading return were reached for MKL with five news categories utilized and two kernels, polynomial and Gaussian, used for each category. Birz and Lott Jr (2011) examined news about GDP growth, unemployment, retail sales, and durable goods and found that the news concerning GDP growth and unemployment significantly affects stock returns. They claim that although having an expected sign, the correlations between stock returns and news about durable goods and retail sales are statistically insignificant, which can be explained by the fact that these variables are less important for investors' expectations of future economic conditions.

As a final example, Schumaker and Chen (2009) explored the discrete stock price prediction using textual representation and statistical machine learning methods on financial news articles partitioned by similar industry and sector groupings. They discovered that stocks partitioned by Sectors were most predictable in measures of Closeness, Mean Squared Error (MSE) score of 0.1954, with a predicted Directional Accuracy of 71.18%, and a Simulated Trading return of 8.50% (compared to 5.62% for the S&P 500 index). In direct comparisons to existing market experts and quantitative mutual funds, the authors' system's trading return of 8.50% outperformed well-known trading experts.

3. Data and methods

In this section, data and methods employed to explore the relationships between financial indices and sentiment scores derived from several financial news segments are described. It begins by examining sentiment scores, which involve the extraction of news content, pre-processing of the text, and the subsequent application of a sentiment analysis model. Then, the causal discovery model is discussed, where the underlying assumptions and the formulation of the model's structure are explored.

3.1 Sentiment analysis

As stated by Wu et al. (2021), online news is considered to be a more accurate source of information when compared to other forms of social media, such as Facebook, X (formerly Twitter), and blogs, due to its greater influence and restrained nature. In this sense, the first step was to write a web crawler for scraping all available news related to five different issues–commodities, stock markets, economy, currencies, and economic indicators–from the web portal investing.com.[3] There is a discussion on whether to use headlines or the entire new text as input to the sentiment model. In the work of Birjali et al. (2021), the authors list levels of sentiment analysis: Document, Sentence, and Aspect levels–where Document-level is the broader one, encompassing the other two. Tsai and Wang (2017) found similar results utilizing sentiment words and all text in their research on financial reports and risk prediction. Considering that each level can contain specific characteristics and contexts, and as a measure of robustness, this research used three so-called levels to determine the sentiment score: at the news title level, at the summary level of the entire news corpus, and at the entire news text level.

Hence, after collecting all the news (their title and corpus), the text corpus was submitted to a summarizing model, which intends to summarize the text

[3]The URL of each news source is the following: Commodities: https://www.investing.com/news/commodities-news; Stock markets: https://www.investing.com/news/stock-market-news; Economic indicators: https://www.investing.com/news/economic-indicators; Economy: https://www.investing.com/news/economy; Currencies: https://www.investing.com/news/forex-news. They have all been accessed from July 6 to July 20, 2023. It would be helpful to describe the methodology of the process employed in assigning each news article to its corresponding topic. However, this information was not accessible on the investing.com website, nor was it publicly disclosed by the editorial staff.

in a maximum length of 1024 tokens. The model used was sshleifer/distilbart-cnn-12-6, which is pre-trained on the CNN/DailyMail dataset, an English-language dataset containing slightly more than 300 thousand unique news articles written by CNN and Daily Mail journalists. Only excessive blank spaces and minor quote typos were removed as pre-processing steps on the news corpus text. [4]

Then, the sentiment model was applied to the news title, the text summary, and the full text. The model used was FinBERT, which is a language model based on Bidirectional Encoder Representations from Transformers (BERT) architecture for the financial domain (Araci, 2019). BERT is essentially a language model consisting of a stack of Transformer encoders. The Global Vectors for the Word Representation (GLoVe) algorithm is used to tokenize the text, i.e., to calculate word representation in a vector space; then, instead of guessing the next word based on the words that came before, BERT "masks" 15% of all tokens. On top of the last encoder layer, a softmax layer over the language predicts the masked tokens. "Next sentence prediction" is the second thing that BERT is trained to do. If two sentences are given to the model, it will tell if they follow each other or not. Token and position embeddings show the order of the input sequence. At the beginning and end of the sequence, two tokens defined [CLS] and [SEP] are added. The [CLS] token is used for all classification tasks, such as figuring out what the next sentence will be (Araci, 2019).

According to BERT's authors (Devlin et al., 2018), it was specifically trained on Wikipedia (2.5B words) and Google's BooksCorpus (800M words). To improve FinBERT's classification performance, the model was further pre-trained in a financial corpus called TRC2-financial, which is a subset of Reuter's TRC2[5]. TRC2-financial includes 46,143 documents with more than 29M words and nearly 400 thousand sentences. Malo et al. (2014)'s Financial PhraseBank is the main sentiment analysis dataset used to fine-tune FinBERT. The Financial Phrasebank is made up of 4845 English sentences picked at random from the LexisNexis collection of financial news. Then, 16 people with backgrounds in business and banking added notes to each sentence. The people who were asked to annotate the sentence were asked to add labels based on how they thought the facts in the sentence might affect the stock price of the company mentioned. The dataset also has information about how much the annotators agree on each line (Araci, 2019).

As the FinBERT's authors pointed out, a test-set accuracy of 97% was

[4] More information on the model can be found at `https://huggingface.co/sshleifer/distilbart-cnn-12-6`, which was accessed on June 6, 2023.

[5] TRC2 consists of 1.8M news articles that were published by Reuters between 2008 and 2010.

achieved in the complete inter-annotator agreement section of the Financial PhraseBank. However, the model occasionally encounters difficulties in accurately determining the relative magnitude of figures, particularly when there is a lack of explicit directional indicators such as the term "increased." Consequently, it may generate predictions that are neutral in nature. The misclassifications of FinBERT exhibit a notable disparity, with 73% of them occurring between positive and neutral labels, in comparison to a mere 5% between negative and positive labels. This observation aligns with both the inter-annotator agreement statistics and logical reasoning. Distinguishing between positive and negative elements is a very straightforward task. However, determining whether a comment reflects a favorable perspective or simply an impartial observation can present a greater level of difficulty (Araci, 2019). To mitigate misclassifications, the model was performed in all three inputs, i.e., in the financial news title, its summary, and full text. A metric was constructed from the classification model where the modal sentiment between the three was assigned so, for example, if the news title reports a positive sentiment, the summary a negative or neutral, and the text also has a positive sentiment, the overall sentiment to that specific news is positive. If they all disagree, then the sentiment is neutral.

Figure 1
Daily sentiment score

Note: The dots represent daily average sentiment scores on the financial topics. The x-axis represents the time period, and the y-axis is the score of the average sentiment.

The maximum length FinBERT allows one to predict the sentiment is 512 tokens. Then, to overcome this problem, specifically when dealing with the summaries and full text that generally exceeds this size, the model was applied in sliding windows of 512 tokens, and an average was taken from then in the end.

Given the data availability and the maximum length reached by most news, the final output from the sentiment model starts on June 9th, 2015, and spans until March 6th, 2023, for all financial news except Economy news, which starts on April 30th, 2018 and ends with the others. Table A1 shows the number of news during the period according to the respective issue. As can be seen, the Economic news started being released only in 2018, requiring a short analysis. The year 2023 has a relatively smaller number of observations since it extends only until July 2023.

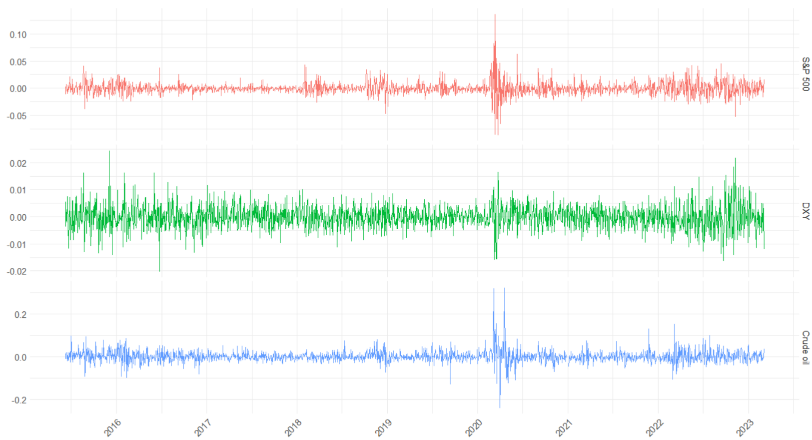
Table A2 shows the number of sentiments per year. The year 2022 is the one with the most negative news among the sample, especially for Stock Market news, with an increase of around 53% when compared with the average in previous periods. This phenomenon may be associated with the steep increase in the interest curve observed in the United States in the first half of 2022. Since then, the US economy has strived to recover from the credit expansion resulting from the COVID-19 pandemic in 2020. In such circumstances, higher interest rates tend to lead to increased investments in fixed-income assets, such as bonds, as opposed to variable-income assets, like equities.

The primary objective of this study is to ascertain the mechanisms via which news influences financial indicators. It is common for financial indices to be reported on a daily basis in the majority of available databases. In the sample period, the average number of news articles per day was 88.4. Therefore, the responses of all news articles were encoded, with negative sentiments assigned a value of -1, neutral sentiments a value of 0, and positive sentiments a value of 1. Next, they were organized based on their sum values, grouped by day and topic. Furthermore, it should be noted that due to the market's operational hours typically limited to 9:30 am to 4 pm during business days, any sentiments derived from news articles published on weekends, holidays, or after the market is closed were carried forward to the subsequent business day. Therefore, the inclusion of a financial news release on a Saturday will contribute to the calculation of the sum for the subsequent Monday, for example. Figure 1 shows the daily average sentiment of each topic. It is possible to note the relatively higher number of negative news ranging from 0 to -1. The Economy and Stock Market news also display this behavior, but with more values ranging from 0 to -0.5, indicating comparatively more positive news than the other topics.

3.2 Financial indices

After summing up the sentiment dataset on a daily basis, financial indices that play a relevant role in the United States of America and the world economy were collected. The first is the S&P 500 index, which is, according to S&P Global, the official S&P website[6], widely recognized as a highly reliable and comprehensive indicator of United States large-cap equities. The index comprises a wide range of 500 major corporations representing various industries within the United States stock market. The index encompasses approximately 80% of the market capitalization of equities in the United States of America and more than 50% of the market capitalization of equities worldwide.

Figure 2
Financial indicators trend



Note: The x-axis represents the sample time period and the y-axis the returns of the financial indicators.

The selection of the second financial measure was the US dollar index (DXY). The composition of the index previously consisted of ten prominent foreign currencies (Harpaz et al., 1990). In current times, it is composed of a basket of six foreign currencies, each with its own respective weightings. These currencies include the Euro, which accounts for 57.6% of the basket;

[6] Available at https://www.spglobal.com/spdji/en/indices/equity/sp-500, accessed October 1st, 2023.

Japanese yen at 13.6%; Sterling Pound at 11.9%; Canadian dollar at 9.1%; Swedish krona at 4.2%, and Swiss franc at 3.6%. The index is determined by calculating the weighted geometric average of the exchange rates between the US dollar and these currencies. To ensure comparability, the resulting values are normalized by an indexing factor of 50.1435. The index functions as a comprehensive metric for evaluating the overall state of the United States economy. According to Chen (2022), traders usually employ it for the purpose of engaging in speculative endeavors, wherein they make predictions regarding variations in the value of the US dollar. Additionally, it can serve as a protective measure against any hazards that may arise from exposure to foreign currencies.[7]

Brent crude oil is the last financial indicator chosen. Crude oil, among many categories of commodities, is extensively traded and exhibits a high volatility behavior. Additionally, there is a connection between the oil market and economic cycles (Hamilton, 1983), affecting several macroeconomic variables such as inflation, GDP, stock market returns, interest rates, and exchange rates (Sharma and Dharmaraja, 2016). Given its importance and together with the other two indicators, different angles in the economy are encompassed.

Figure 2 illustrates the return[8] trend of each indicator. It is noteworthy that all three indicators experienced a period of heightened stress in the year 2020, specifically because of the global COVID-19 pandemic. Furthermore, from the start of 2022 until the end of the sample period, July 2023, there has been an observable increase in volatility when compared to pre-pandemic times, particularly in relation to the S&P 500 index. Besides the financial indicators, the Dow Jones Industrial Average and the volatility index (VIX) are used as covariates in the discovery model. Daily values for all financial indicators were gathered from the Yahoo Finance website throughout the time span covered by the financial news dataset.

3.3 Causal discovery model

The causal discovery model was based on the paper “Causal inference for time series” by Runge et al. (2023). It was implemented with the Python Tigramite package. As the authors indicated, causal inference offers a comprehensive framework that combines statistical and machine-learning techniques to address causal inquiries using data. Its utilization allows for the formulation of research questions in a causal manner while also providing a

[7] For a full understanding of the methodology employed on the US index dollar, access https://www.ice.com/publicdocs/data/ICE_FX_Indexes_Methodology.pdf.

[8] The returns are expressed by $Returns = (x_{t-1} - x_t)/x_t$.

clear exposition of the underlying assumptions employed to address them. In other words, causal discovery involves the qualitative reconstruction of connections within a causal graph, either in its entirety or specifically focusing on the causes of a particular target variable. This reconstruction may incorporate elements such as temporal lags and bi-directed links to indicate latent confounding.

Financial markets are complex systems influenced by a multitude of known and unknown variables, and many of these variables are unobservable or difficult to quantify directly. As examples of confounding variables, there are the investor's behavior, sentiment, and psychological biases; government policies, central bank decisions, and regulatory changes; information asymmetry; unpredictable and rare events; and random noise and unexpected shocks. These factors are essentially unobservable and can lead to short-term market volatility and fluctuations. Hence, the model used in this work is the Latent-PCMCI algorithm, which allows the presence of unobserved factors. As the model contains a specific notation, first, a brief introduction is given, and then the model is discussed.

3.3.1 LPCMCI background

The LPCMCI model requires one to make assumptions about the data-generation process of the variables of interest. They are discussed throughout this Section. The first requirement is the absence of selection bias, which occurs when the process of selection data for our analysis systematically influences the causal relationships that we aim to uncover. Selection bias can arise from various factors, as discussed by Reiser (2022). These factors include (i) non-random sampling, which occurs when the sample does not accurately represent the characteristics of the population; (ii) missing data, where there is a systematic pattern in the covariates resulting in missing values for a particular subgroup of the population; (iii) survivorship bias, which arises when only processes that have endured over time are observed. We believe that such concerns are mitigated in our empirical approach since we are using news articles from an international authority source in finance and economics (investing.com), which covers facts that directly affect financial and economic variables worldwide. Therefore, our economic and financial variables are not limited to the macroeconomic performance of a specific subgroup of countries (e.g., developed countries) or a continent, reducing concerns with non-random sampling. Additionally, missing data is not a concern for our empirical specification. Ultimately, we analyze time series data instead of panel-data format, thereby mitigating concerns related to survivorship bias.

Following the methodology proposed by Gerhardus and Runge (2020) and the Tigramite documentation, the focus is to learn the causal structure underlying complex dynamical systems, hence the time series $\mathbf{V}_t = (V_t^1, \ldots, V_t^N)$ is assumed to follow a structural causal process of the form

$$V_t^j = f_j\left(\mathscr{P}(V_t^j), \eta_t^j\right) \tag{1}$$

where f_j is some arbitrary measurable function that also depends non-trivially on all its arguments; η_t^j is the dynamical noise variable which is jointly independent, i.e., $(i \neq j)$ and $(t' \neq t)$; and the set $\mathscr{P}(V_t^j) \subset \mathbf{V}_{t+1}^- = (\mathbf{V}_t, \mathbf{V}_{t-1}, \ldots) \setminus \{V_t^j\}$ define the causal parents of V_t^j. In this sense, the value of V_t^j is said to be determined by the values of the variables in $\mathscr{P}(V_t^j)$ and the value of the dynamical noise η_t^j, and hence, it is claimed that the equation has causal significance.

Moreover, it is required that $V_{t-\tau}^i \in \mathscr{P}(V_t^j)$ if and only if $V_{t-\tau-\Delta t}^i \in \mathscr{P}(V_{t-\Delta t}^j)$ which represents the assumption of *causal stationarity*, where there are no cyclic causal relationships. To ensure stationarity, an Augmented Dickey–Fuller (ADF) unit root tests were conducted in all the variables used in the causal discovery model. The null hypothesis provided by the test assumes that the time series exhibits a unit root, which implies that it is non-stationary. In practical terms, this means that the data have a stochastic or random trend, and its statistical properties do not remain constant over time.[9] If a variable is found to be non-stationary, a first-order difference is taken from it, and the ADF test is applied again to ensure stationarity, where no further differencing is needed.

The objective of the LPCMCI is to acquire knowledge about the time series Directed Partial Ancestral Graph (DPAG) of the data-generating structural causal process. The DPAG is a graphical representation that captures partial information regarding the causal ancestral relationships solely among the observable variables. The structural causal relationships defined by a structural causal process can be effectively illustrated using a directed acyclic graph (DAG) $\mathscr{G}$ that has one node per variable V_t^j and a link $V_{t-\tau}^i \to V_t^j$ if and only if $V_{t-\tau}^i \to V_t^j$.[10]

In the LPCMCI model, the set of component time series $V^1, \ldots, V^N$ can be divided into a group of observed time series $X^1, \ldots, X^{N_X}$ with $N_X \geqslant 1$ and a group of unobserved time series $L^1, \ldots, L^{N_L}$ with $N_L \geqslant 0$ and $N = N_X +$

[9] For more details on the ADF unit root test, refer to Dickey and Fuller (1979, 1981).

[10] As mentioned in the Tigramite package documentation, the acyclicity of the system is a direct consequence of the absence of cyclic causal links, as assumed. Furthermore, the repeated temporal structure of the system can be attributed to its causal stationarity.

N_L. In order to illustrate the causal relationships of the main process using a graph that includes only the observed variables, the LPCMCI utilizes Directed Maximal Ancestral Graphs (DMAGs). The idea is that to a given DAG $\mathscr{G}$ with a given subset of unobserved variables, it is possible to associate a unique DMAG $\mathscr{M}(\mathscr{G})$ over the observed variables.

The $\mathscr{G}$ holds causal significance as it represents causal relationships through its directed edges. Similarly, the associated $\mathscr{M}(\mathscr{G})$ also conveys causal meaning. In summary, the DMAG $\mathscr{M}(\mathscr{G})$ serves as a representation of the causal ancestral links relative to the underlying data-generating process. However, the existence or absence of an edge between two variables does not have a straightforward causal interpretation. Considering that only a finite number of time steps can be observed on a time series leads to the choice of an observed time frame $[t-\tau_{\max},t]$, where $\tau_{\max} \geqslant 0$ is referred as the maximum considered time lag. Furthermore, the recurring arrangement of the underlying $\mathscr{G}$ can also be utilized to enforce a repeating structure on the corresponding $\mathscr{M}(\mathscr{G})$, which is subsequently extended to time steps beyond the observed time window.

The next step is to learn time series DMAGs from observations of the observed time series. LPCMCI accomplishes this within the constraint-based approach to causal discovery that employs (conditional) data interdependencies. However, it should be noted that this problem is under-determined since different DMAGs can result in the exact same set of (conditional) independencies, a concept known as Markov equivalence. Therefore, it is generally not feasible to uniquely discover the DMAG $\mathscr{M}(\mathscr{G})$ that accurately represents the causal ancestral relationships of the underlying data-generating process. Instead, what can be obtained is a collection of potential time series DMAGs $\mathscr{M}_1,\ldots,\mathscr{M}_m$ which includes $\mathscr{M}(\mathscr{G})$, and forms the Markov equivalence class of $\mathscr{M}(\mathscr{G})$. Hence, only features that are shared by every member of the Markov equivalence class of $\mathscr{M}(\mathscr{G})$ can be learned. These common features can, therefore, be expressed using directed partial ancestral graphs (DPAGs).

DPAG $\mathscr{P}(\mathscr{G})$ main idea is that all members of the Markov equivalence class agree on adjacencies. In other words, X and Y are connected by an edge in $\mathscr{M}(\mathscr{G})$ if and only if they are connected by an edge in all members of the Markov equivalence class. However, the members can have different edge types. If $X \rightarrow Y$ is in $\mathscr{M}(\mathscr{G})$ then there could be a member in the Markov equivalence class of $\mathscr{M}(\mathscr{G})$ in which $X \leftarrow Y$ or $X \leftrightarrow Y$. These ambiguities are represented explicitly by $X {\circ\!\!\rightarrow} Y$ and non-directed edges $X \circ\!\!-\!\!\circ Y$.

The LPCMCI model requires the Markov equivalence class together with the faithfulness assumption, which states that there are no variables determining which measurements are included or excluded from the data sample. Conditional independencies in the observed distribution can happen by chance, which means that faithfulness can be violated. However, as sample sizes get bigger, the chance of violation becomes less possible (Reiser, 2022). Fundamentally, the correspondence between both conditions implies that conditional independencies in the probability distribution are due to the causal relationships represented by the graph (Eberhardt, 2017; Jiang and Shimizu, 2023).

Because the $\mathcal{M}(\mathcal{G})$ causal meaning in the sense that its edges represent causal ancestral links, the related DPAG $\mathcal{P}(\mathcal{G})$ has causal meaning as well: (a) $X \rightarrow Y$ says the same as in $\mathcal{M}(\mathcal{G})$, X is a (potentially indirect) cause of Y, and Y does not cause X; (b) $X \leftrightarrow Y$ says the same as in $\mathcal{M}(\mathcal{G})$, X does not cause Y, Y does not cause X, and X and Y are subject to unobserved confounding, i.e., there is an unobserved variable Z that causes both X and Y; (c) $X {\circ\!\!\rightarrow} Y$ means that X may or may not cause Y, and Y does not cause X; (d) $X {\circ\!\!-\!\!\circ} Y$ means that X may or may not cause Y, and Y may or may not cause X.

Hence, the DPAG $\mathcal{P}(\mathcal{G})$ serves as a representation that captures a partial knowledge of the causal ancestral relationships inside the data-generating process. Regarding $\mathcal{M}(\mathcal{G})$, it is important to note that the presence or absence of an edge connecting two variables does not possess a straightforward causal interpretation.

3.3.2 LPCMCI algorithm

As Gerhardus and Runge (2020) explain, the fundamental basis of the LPCMCI is the FCI algorithm. The central feature of the LPCMCI is the observation that strong autocorrelations tend to reduce the effect sizes of (conditional) independence tests, thereby reducing the statistical power of these tests and, thus, degrading the algorithm's overall statistical performance. The aim is to alleviate this problem by conditioning the autocorrelation away (at least partially). This is achieved by extending the standard conditioning sets $\mathcal{S}$ of (conditional) independence tests with so-called default conditions $\mathcal{S}_{def}(X^i_{t-\tau}, X^j_t)$, schematically

$$\text{Test whether } X^i_{t-\tau} \perp X^j_t \mid \mathcal{S} \quad \longrightarrow \quad \text{Test whether } X^i_{t-\tau} \perp X^j_t \mid \mathcal{S} \cup \mathcal{S}_{def}(X^i_{t-\tau}, X^j_t)\,. \tag{2}$$

In order to learn causal ancestor links while still testing for additional (conditional) dependencies, the LPCMCI makes use of a novel set of orientation rules that extends those of the FCI method. All subsequent tests will utilize the identified causal ancestors as their default conditions $\mathscr{S}_{def}(X^i_{t-\tau}, X^j_t)$. The algorithm alternates between running (conditional) independence tests and applying the orientation rules.

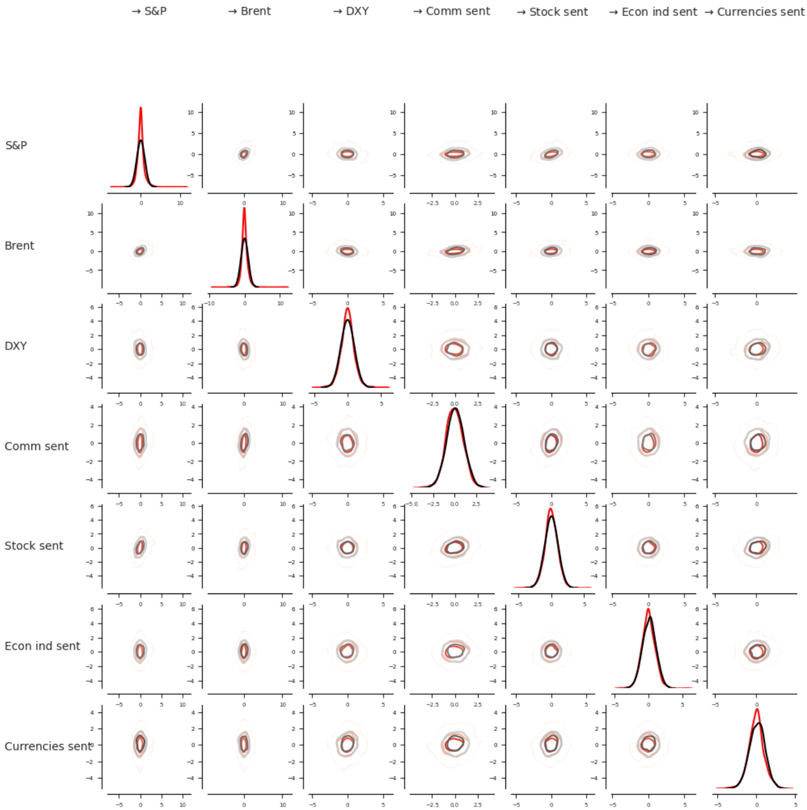
3.3.3 Conditional independence tests: Gaussian Process Regression

To uncover the causal relationships between the variables, the model employs conditional independence tests. There are several different methods used to incorporate them. The chosen approach is based on Gaussian process regression and a distance correlation on the residuals. The Gaussian processes model is a probabilistic supervised machine learning framework that has been widely used for regression and classification tasks (Wang, 2020). Specifically, Gaussian processes are mathematically equivalent to many well-known models, including Bayesian linear models, spline models, large neural networks (under suitable conditions), and are closely related to others, such as support vector machines (Williams and Rasmussen, 2006).

In problems where it is required to assess unknown functions that map inputs to outputs, often the shape of the underlying function is unknown, the function might be difficult to evaluate analytically, or other requirements such as design costs might complicate the process of information acquisition. Gaussian process regression is a powerful, non-parametric Bayesian approach towards this type of regression problem (Schulz et al., 2018). As specified by Silva Filho et al. (2014), multivariate distributions are needed to explain data with varied dispersion patterns, including asymmetry, heavy tails, and other behaviors commonly found in financial returns. Furthermore, standard measures of dependence, such as the Pearson correlation coefficient, may exhibit limitations in capturing the non-linear structures of dependence present in bivariate data, particularly when the variables involved are non-linear (Nguyen et al., 2020). Figure 3 depicts the density plots of the data distribution combinations. In general, all plots hardly resemble a linear behavior, confirming the necessity of a non-parametric approach. A full description of the Gaussian process regression methodology can be found in Williams and Rasmussen (2006), Schulz et al. (2018), and Wang (2020). In this article, only an overview and the intuition behind the method are discussed.

As a starting point, regression analysis involves fitting a function to represent certain observed data points and using the function to predict new data points. However, there can be an endless number of functions that fit a given collection of observed data points. In order to perform a regression, the Gaussian process defines a distribution over this infinite set of functions (Wang,

Figure 3
Joint and marginal densities plot



2020). To provide more context, imagine a set of Gaussian vectors[11] connected by random selected sample points of each vector as lines. The result would be similar to several fraught functions, i.e., they would not be smooth enough for regression tasks (Wang, 2020).

The smoothness can be achieved by defining covariance functions or kernels, that reflect the ability to express prior knowledge of the form of the

[11] A Gaussian vector has $P_X(x) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ as its probability density function (PDF) where μ is the mean, σ^2 is the variance, X are random variables, and x is the real argument (Wang, 2020).

function to be modeled (Wang, 2020). The kernel used to determine the conditional independence test is the radial basis function (RBF) kernel, which provides an expressive kernel to model smooth and stationary functions as it has two parameters – λ, called the length-scale, and σ_f^2, the signal variance – that can be varied to increase or reduce a prior correlation between data points and hence, the variability of the resulting function Schulz et al. (2018). In this sense, the prior distribution represents the expected outputs of f over inputs x without observing any data. Once there are data to work with, it is possible to eliminate the countless functions that would not fit and keep only the ones that do. The observed data have been included in the prior, creating the posterior. This process keeps consecutively with the current posterior as prior and new observed data to obtain a new posterior. Finally, the mean function calculated by the posterior distribution of possible functions is the function used for regression predictions (Wang, 2020).

4. Results

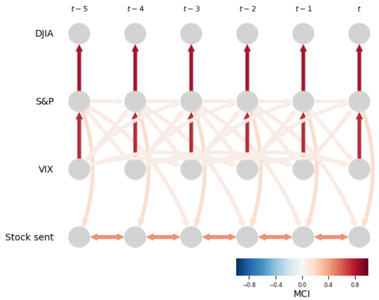
4.1 News and financial variables effects

Before turning the analysis to the causal discovery model, three parameters must be set: the $\tau_{\max}$ which indicates maximum time lag considered, i.e., the number of past periods used to investigate the lagged dependency; the significance level α_{PC} of the individual (conditional) independence tests, where higher values tend to make the estimated graph denser; and the number of preliminary iterations k of the LPCMCI algorithm[12]. The $\tau_{\max}$ was set to 5 to encompass a full working week. Although news and financial indicators can relate in a greater time window (Walker, 2016), the computational cost is significantly increased as the number of time periods rises. α_{PC} was set to 0.01 to avoid misclassifications, and $k = 2$ is a conservative value. As mentioned before, all variables are stationary according to the Dickey-Fuller unit root test, but only the sentiment variables were not stationary in level, requiring a first-order differencing. Another point of attention is that all the variables were standardized to the same scale with the z-score formula $z = \frac{(x-\mu)}{\sigma}$, where x is the data values, μ its mean, and σ its standard deviation.

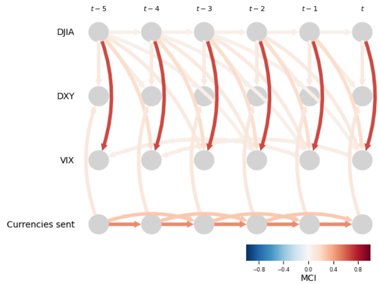
Figure 4 presents the article's main results. The causal discovery model was applied to each financial variable in combination with each news topic sentiment. A complete exposition of the results can be requested from the

[12] As shown in the Tigramite documentation, not including preliminary iterations ($k = 0$) can lead to erroneous links between the variables and their lags, which is reflected by a low effect size as a consequence of the autocorrelation of the time series. Setting k to greater values addresses the problem by restoring all removed edges after the preliminary phase.

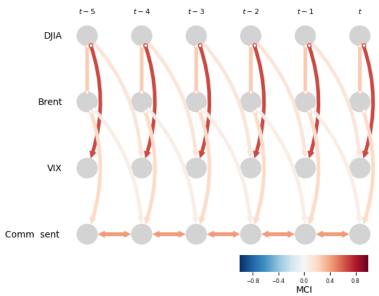
Figure 4
Causal discovery PDAG on financial indicators and news sentiment topics



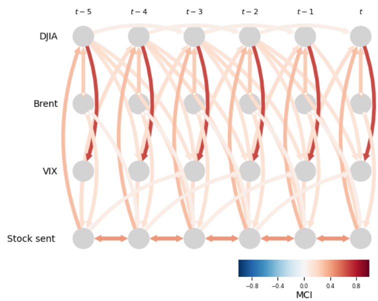
(a) S&P 500 and Stock Market news sentiment



(b) DXY and Currencies news sentiment



(c) Brent and Commodities news sentiment



(d) Brent and Stock Market news sentiment

authors. In Figure 4 (a) a contemporaneous and a two-day lag relationship is seen from the S&P index to the Stock Market news sentiment (Stock sent). As highlighted in Subsection 3.3, the one-way arrow from the S&P to the Stock sentiment indicates that the S&P is a (potentially indirect) cause of the Stock Market news sentiment. The color of an edge or node represents the strength of the cross- or auto-dependency, respectively (Reiser, 2022). This result suggests that S&P's volatility today is reflected in the news sentiment also today and after tomorrow. Conversely, in 4 (b), the causal link is found between the Currencies sentiment news (Currencies sent) to the DXY index on the same day, without any lagged effect. Figure 4 (c) illustrates a relation between Brent crude oil and the Commodities news sentiment (Comm sent). However, the link is of shape $\circ\!\!\rightarrow$, which means two results: first, Brent crude oil's volatility may or may not cause Commodities news sentiments; and second, the sentiments do not have an effect on Brent's returns. In a similar way, 4 (d) depicts the same interaction between the Brent crude oil and the Stock Market news sentiment, i.e., ($\text{Brent}_t \circ\!\!\rightarrow$ Stock sent_t) and ($\text{Brent}_{t-1} \circ\!\!\rightarrow$ Stock sent_t). There is, though, in this setting, a current causal link between the Stock Market news sentiment to the Dow Jones Industrial Average (DJIA) index, and a $\text{DJIA}_{t-2} \rightarrow$ Stock sent_t.

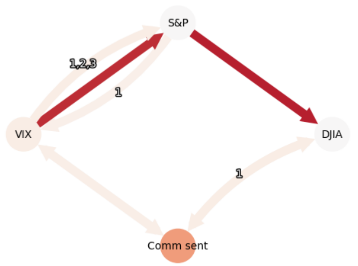
4.2 Secondary effects

Figure 5 shows the indirect effects of the interactions between the variables. In order to differentiate from the main results and to facilitate the view, they are depicted in a different type of graph where the lag interactions are defined by a number in the middle of the arrows, which indicates the lag order. Figure 5 (a) and (b) present the LPCMCI structure used to model the variable S&P 500 index. In (a), it is possible to note an interaction between VIX and DJIA with a day lag by a two-head arrow. This arrow type suggests that any of the variables cause each other directly. Instead, however, it means that an unobserved third variable is causing both. The same occurs in Figure 5 (b) with VIX. Figures 5 (c) and (d) are constructed following the DXY and Brent structures, respectively. In (c), there is the same indication of a confounding variable affecting Stock news sentiment and VIX. In turn, there is a potentially indirect effect of Stock Market news sentiment towards DJIA, in level and as well with one lag. In (d), there is probably an indirect effect of VIC on Currencies news sentiment.

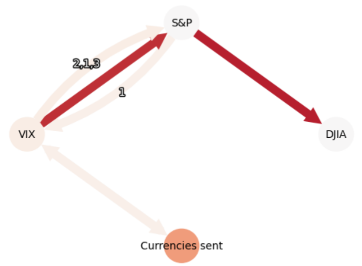
4.3 Robustness checks

Probably the most naive approach to infer a relationship between sentiments and financial variables would be a correlation analysis. Figure 6 displays a correlation plot with all variables of interest and each financial news category. Starting on the first row and excluding the non-significant results at $\rho < 0.05$, it is possible to note a correlation of 0.2 between the Commodities news sentiment and the Brent crude oil variable. Although it is a small value, it shows that the variables were connected during the analysis period. The other statistically significant correlation among the financial variables is with the S&P 500 index, showing a small degree of correlation. The sentiments from the Stock Market news are linked with the S&P indicator with a level of 0.33, the highest registered among the variables; and with Brent. As pointed out previously, the lack of the methodological criterion used in financial news segmentation inhibits a deeper analysis. However, Economy and Economic Indicators news topics seem to be similar, showing no statistically significant correlation with the financial variables. Finally, the Currency news sentiment is only related to the DXY.

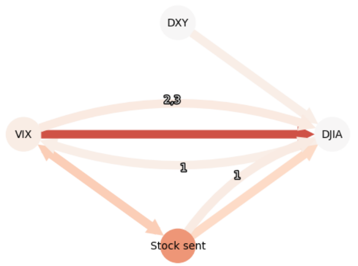
Figure 5
Causal discovery PDAG on financial indicators and news sentiment topics – secondary results



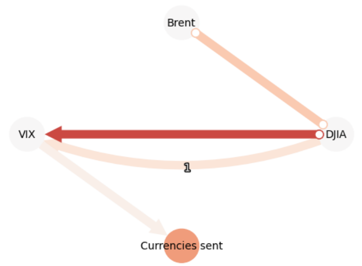
(a) S&P 500 setting and Commodities news sentiment agreed



(b) S&P 500 setting and Currencies news sentiment agreed



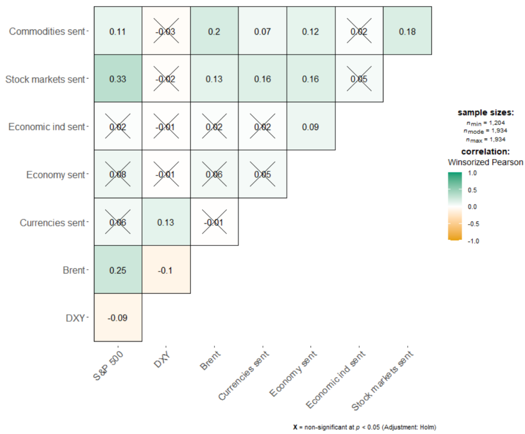
(c) DXY setting and Stock Market news sentiment agreed



(d) Brent setting and Currencies news sentiment agreed

As a measure of robustness, a second metric for the sentiment score was used. As explained in Section 3.1, the first one utilized the modal sentiment between the three, so, for example, if the news title reports a positive sentiment, the summary a negative or neutral one, and the text a positive sentiment, the overall sentiment on that specific news is positive. If they all disagree, then the sentiment is neutral. The second metric is more restrictive. Only the sentiments where all results were the same were used. As the sentiment model could not be evaluated against a sample with sentiments attributed by financial specialists, for example, this restricted dataset aims to be more reliable than the original one. On the downside, however, fewer observations are computed on the intraday basis. Nonetheless, it is important to note that, although the daily sentiment dataset becomes smaller, the number of observations on the merged dataset with the financial indices is the same. This occurs because

Figure 6
Correlation plot

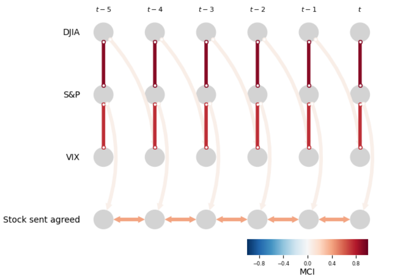


the number of news per day decreases, but all days have at least one estimated sentiment where all the inputs return the same result. In this setting, the average number of news per day is 38.9.

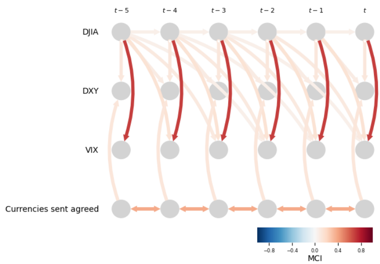
Figure 7 shows the most prominent results. The term “agreed” is used to highlight the difference between both datasets. Analyzing the relationship between the S&P 500 index and the Stock Market news sentiment in Figure 7 (a), it is possible to note a difference from the previous result. Only a current effect of the S&P index is found towards the sentiment variable and with a left-circled arrow ($\circ\!\!\rightarrow$). This decreases the confidence level of the effect since it could or could not have a causal relationship between the variables. In Figure 7 (b) and (c) the same causal relationship between the variables is noted, i.e., Currencies news sentiment$_t$ $\rightarrow$ DXY$_t$, and Brent$_t$$\circ\!\!\rightarrow$ Commodities news sentiment$_t$ and Brent$_{t-1}$$\circ\!\!\rightarrow$ Commodities news sentiment$_t$, respectively. Figure 7 (d) shows the opposite effect between the variables, i.e., the Stock sentiment affects Brent crude oil returns with a straight arrow. The last graph in Figure 7 (e) provides a result not found in the complete dataset. A relation of three lags is seen from the Economy news sentiment towards Brent crude oil, represented by the left circled arrow, which means that there can or cannot exist a causal effect between the variables.

Following a considerably popular approach in the literature (e.g., Hiemstra and Jones, 1994; Wilms et al., 2016; Baum et al., 2021), the Granger

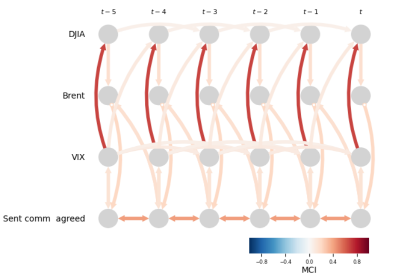
Figure 7
Causal discovery PDAG on financial indicators and news sentiment topics agreed



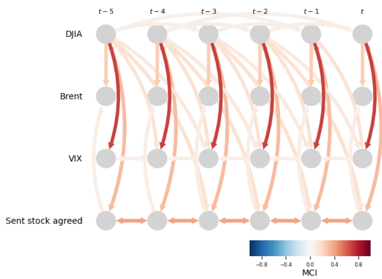
(a) S&P 500 and Stock Market news sentiment agreed



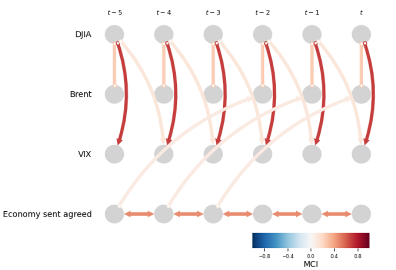
(b) DXY and Currencies news sentiment agreed



(c) Brent and Commodities news sentiment agreed



(d) Brent and Stock Market news sentiment agreed



(e) Brent and Economy news sentiment agreed

Causality test[13] was performed in the variables in order to compare the re-

[13]The Granger Causality test (Granger, 1969) posits that if variable A exhibits Granger causality on variable B, then the historical values of A should contribute to the prediction of future values

sults. The test does not consider contemporaneous effects, i.e., at lag 0. However, the correlation plot in Figure 6 gives an idea of the current relationship between the variables. Table A3 highlights only the statistically significant results. It is worth noting that several lag interactions between the variables are captured in the test, including the ones in bold that are also depicted in the LPCMCI model.

5. Discussion

The main objective of this work was to systematically investigate the intricate relationship between different financial news topics and relevant financial indicators, specifically focusing on the S&P 500 index, dollar index (DXY), and Brent crude oil. The results suggest a complex relationship between the financial variables and the financial news sentiments. For instance, the S&P 500 index, which presents a level of correlation with the Stock Market news, can be seen as a source of information for the Stock Market news content. This kind of relationship is not exclusive to these variables. Brent crude oil and Commodities news also appeared to have a similar behavior.

In the opposite direction, Currencies news sentiment is impacting DXY volatility. These results suggest an ambiguous effect between news and financial variables, as pointed out by Walker (2016). Nonetheless, this behavior is not extended to all sentiment topics. Economy and Economic indicators did not show any correlation with the financial variables. Thus, it seems that the news sentiment is indeed more relevant depending on the financial sector.

Another point of interest is the lag dependency observed in the financial variables towards the news sentiments. In the case of Brent leading to Commodities news sentiment variations, the observed lag was one day, which is to be expected, since the news sentiments for the subsequent day start at the end of the trading time, i.e., after the market's operational period, one can expect to see news regarding the market performance. In the case of S&P 500 returns toward Stock Market news sentiments, the observed lag is two days. It is possible that as the S&P index comprises a large number of companies, processing and interpreting the impact of S&P 500 returns on the overall stock market may take more time. It could require a period of data analysis, research, and evaluation before a news sentiment is expressed and disseminated. The lag may also result from cumulative effects, where market movements may need

of B, surpassing the predictive capability only based on previous values of B. It is important to acknowledge, however, that the test only considers linear relationships and that this does not show causation in a deterministic manner. Instead, it presents statistical data that support a cause-and-effect connection, indicating that alterations in one variable are linked to later alterations in another.

to accumulate over a short period before they are deemed significant enough to influence sentiment substantially. This behavior might not be seen in the case of Brent as it is a more targeted market.

Regarding the impacts on sentiments felt on the same operational day, one may argue that during trading hours, market volatility can be heightened. Sudden price swings and significant events can elicit immediate media responses, which are translated as news sentiments. In this case, the opposite also holds since high-frequency traders and algorithmic trading systems often operate in real-time, making quick decisions based on breaking news as it unfolds (Boubaker et al., 2021).

Depending on the configuration of the model employed to assess the relationship between the variables, news sentiment variables showed a connection with VIX and DJIA. In general, this connection is due to a third non-observable variable identified by the LPCMCI model. External events, such as global economic crises, geopolitical tensions, or unexpected economic releases, can impact both news sentiment and financial market behavior. Such exogenous events can trigger changes in news reporting and simultaneously affect market volatility. Additionally, news sentiments often reflect market expectations and reactions to anticipated events. A non-observable third variable might capture the underlying market sentiment, including investor expectations, uncertainty, and a risk appetite, which in turn influences both news content and market indicators like VIX and DJIA. Finally, behavioral finance principles, such as herding behavior, could explain the simultaneous changes in both news sentiment and financial variables. Investors may react to market fluctuations in a manner that influences news sentiment, and vice versa, restating the variables' mutual effect.

The robustness tests accredited the results found in the main analysis. The correlation and Granger causality tests showed that, indeed, there is a connection between the financial variables and the news sentiments. It is important to note, however, that the methodology employed by those tests does not consider non-linearity and other complexities intrinsic to financial variables. The analysis with the dataset where only the agreed sentiments between the three inputs (news headline, corpus, and summary) of the sentiment model were used returned similar results, with the exception of the relationship between Brent and Stock news sentiments. On the "agreed" dataset, the relation is uni-directed from the Stock news sentiment towards Brent returns. As the first results do not possess a high confidence level given by the circled arrow, it probably captured an incorrect connection, which was corrected on the restricted dataset. Furthermore, the model also captured a three-day lag from the Economy news sentiment toward Brent. Economy news probably encompasses macroeconomic signs, which tend to be well anticipated by the market (Ehrmann and Fratzscher, 2005). If the actual data align with these expectations, it may not lead to an immediate market response. However, it is important to note that the length of the Economy news sentiment period starts only on April 30, 2018, which is considerably shorter when compared with the other topics.

On the subject of methodology, the LPCMCI model is considerably innovative in analysing sentiments and financial variables. Besides its capability of causal discovering considering lag dependencies and structural form, the independence tests based on Gaussian process regressions provide a flexible approach to capture complex relationships between the variables by its non-parametric specification. In this manner, it is expected to effectively model all characteristics present in the data. Moreover, the news sentiment data highlight the benefits of combining NLP and LLM with the financial and economic literature.

6. Conclusion

This work addressed the relationship between the financial news sentiment and financial indicators. A news sentiment database was constructed through the extraction of financial news from different topics from the investing.com web portal. The news was then submitted to the FinBERT model to draw out its respective sentiments. The news dataset was merged with financial indicators, namely the S&P 500 index, the dollar index (DXY), and Brent crude oil, as they represent a diverse range of sectors within the United States economy. Then, the LPCMCI model was implemented to perform causal discovery on the variables of interest.

The main results showed that there is a bi-directed effect between news sentiments and financial variables, as pointed out by Walker (2016). In this sense, the S&P 500 index can be seen as a source of information for Stock Market news content. In the case of the news sentiment on the Commodities subject, a direct contemporaneous and lagged effect is observed towards Brent returns. However, in the opposite direction, news sentiments regarding the Currencies subject appear as a driver of DXY returns. It was also noted that different topics interact in different ways with each financial indicator. In this sense, news related to the Stock Market, Commodities, and Currencies are more linked to the S&P 500, Brent crude oil, and DXY, respectively. Moreover, the results were accredited by popular statistical tests in the literature.

It is important to acknowledge some limitations. For instance, the causal relationship considered in this work is mainly statistical and does not encompass an economic causal relation. Furthermore, the methodological approach used by investing.com on their news topic segmentation was not available, which is a piece of relevant information. Finally, the LPCMCI model requires assumptions such as the absence of variable selection bias. Although robustness tests aimed to mitigate this problem, it cannot be completely ignored.

This work contributes significantly to the fields of financial and economic research as well as to the existing literature in computer science, given the close relationship between the methods used, such as NLP, LLM, and machine learning techniques, and the field of computer sciences. First, it elucidates the intricate connection between the financial news sentiment and significant financial indicators, covering many sectors of the US economy (including the stock market, commodities market, and foreign currency market). Second, the research makes use of cutting-edge methods, including the Latent PCMCI model and Gaussian process regression, in addition to sentiment analysis. These methodological contributions increase the toolkit available to researchers and provide a precedent for future inquiries in this field. Finally, this work evaluates the outcomes in light of conventional statistical procedures, including correlation and Granger causality tests. Future works may quantify the size of the causal effect found among the sentiment and financial variables.

Acknowledgements

This research was funded by LARSyS (Projeto—UIDB/50009/2020). M. Vinicius Santos gratefully acknowledges the financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. Thiago C. Silva (Grant no. 302703/2022-5) acknowledges financial support from the CNPq Foundation.

References

Alanyali, M., Moat, H. S. and Preis, T. (2013). Quantifying the relationship between financial news and the stock market, *Scientific reports* **3**(1): 3578.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models, *CoRR* **abs/1908.10063**.
URL: http://arxiv.org/abs/1908.10063

Balcilar, M., Bekiros, S. and Gupta, R. (2017). The role of news-based uncertainty indices in predicting oil markets: a hybrid nonparametric quantile causality method, *Empirical Economics* **53**: 879–889.

Baum, C. F., Hurn, S. and Otero, J. (2021). The dynamics of us industrial production: A time-varying granger causality perspective, *Econometrics and Statistics* .

Birjali, M., Kasri, M. and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends, *Knowledge-Based Systems* **226**: 107134.

Birz, G. and Lott Jr, J. R. (2011). The effect of macroeconomic news on stock returns: New evidence from newspaper coverage, *Journal of Banking & Finance* **35**(11): 2791–2800.

Blasco, N., Corredor, P., Del Rio, C. and Santamaría, R. (2005). Bad news and dow jones make the spanish stocks go round, *European Journal of Operational Research* **163**(1): 253–275.

Boubaker, S., Liu, Z. and Zhai, L. (2021). Big data, news diversity and financial market crash, *Technological Forecasting and Social Change* **168**: 120755.

Chen, J. (2022). What Is the U.S. Dollar Index (USDX) and How to Trade It — Investopedia, `https://www.investopedia.com/terms/u/usdx.asp`. [Accessed October 6, 2023].

Chen, Y. and Hao, Y. (2018). Integrating principle component analysis and weighted support vector machine for stock trading signals prediction, *Neurocomputing* **321**: 381–402.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American statistical association* **74**(366a): 427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica: journal of the Econometric Society* pp. 1057–1072.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery, *International Journal of Data Science and Analytics* **3**: 81–91.

Ehrmann, M. and Fratzscher, M. (2005). Transparency, disclosure and the federal reserve, *Available at SSRN 677434* .

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work, *The journal of Finance* **25**(2): 383–417.

Feuerriegel, S. and Gordon, J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions, *European Journal of Operational Research* **272**(1): 162–175.

Feuerriegel, S., Ratku, A. and Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation, *2016 49th Hawaii international conference on system sciences (HICSS)*, IEEE, pp. 1072–1081.

Gerhardus, A. and Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders, *Advances in Neural Information Processing Systems* **33**: 12615–12625.

Glymour, C., Zhang, K. and Spirtes, P. (2019). Review of causal discovery methods based on graphical models, *Frontiers in genetics* **10**: 524.

Gong, X., Guan, K. and Chen, Q. (2022). The role of textual analysis in oil futures price forecasting based on machine learning approach, *Journal of Futures Markets* **42**(10): 1987–2017.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: journal of the Econometric Society* pp. 424–438.

Hamilton, J. D. (1983). Oil and the macroeconomy since world war ii, *Journal of political economy* **91**(2): 228–248.

Harpaz, G., Krull, S. and Yagil, J. (1990). The efficiency of the us dollar index futures market, *The Journal of Futures Markets (1986-1998)* **10**(5): 469.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation, *The Journal of Finance* **49**(5): 1639–1664.

Hisano, R., Sornette, D., Mizuno, T., Ohnishi, T. and Watanabe, T. (2013). High quality topic extraction from business news explains abnormal financial market volatility, *PloS one* **8**(6): e64846.

Jiang, Y. and Shimizu, S. (2023). Linkages among the foreign exchange, stock, and bond markets in japan and the united states, *arXiv preprint arXiv:2310.16841* .

Kalyani, J., Bharathi, P., Jyothi, P. et al. (2016). Stock trend prediction using news sentiment analysis, *arXiv preprint arXiv:1607.01958* .

Ko, C.-R. and Chang, H.-T. (2021). Lstm-based sentiment analysis for stock price forecast, *PeerJ Computer Science* **7**: e408.

Kordonis, J., Symeonidis, S. and Arampatzis, A. (2016). Stock price forecasting via sentiment analysis on twitter, *Proceedings of the 20th pan-hellenic conference on informatics*, pp. 1–6.

Li, Z., Huang, Z. and Failler, P. (2022). Dynamic correlation between crude oil price and investor sentiment in china: Heterogeneous and asymmetric effect, *Energies* **15**(3): 687.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts, *Journal of the Association for Information Science and Technology* **65**(4): 782–796.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T. and Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers, *IEEE access* **8**: 131662–131682.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis, *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)*, IEEE, pp. 205–208.

Nguyen, Q. N., Aboura, S., Chevallier, J., Zhang, L. and Zhu, B. (2020). Local gaussian correlations in financial and commodity markets, *European journal of operational research* **285**(1): 306–323.

Oberlechner, T. and Hocking, S. (2004). Information sources, news, and rumors in financial markets: Insights into the foreign exchange market, *Journal of economic psychology* **25**(3): 407–424.

Reiser, C. (2022). Observational and interventional causal learning for regret-minimizing control, *arXiv preprint arXiv:2212.02435* .

Runge, J., Gerhardus, A., Varando, G., Eyring, V. and Camps-Valls, G. (2023). Causal inference for time series, *Nature Reviews Earth & Environment* **4**(7): 487–505.

Schulz, E., Speekenbrink, M. and Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions, *Journal of Mathematical Psychology* **85**: 1–16.

Schumaker, R. P. and Chen, H. (2009). A quantitative stock prediction system based on financial news, *Information Processing & Management* **45**(5): 571–583.

Sharma, H. and Dharmaraja, S. (2016). Effect of outliers on volatility forecasting and v alue at r isk estimation in crude oil markets, *OPEC Energy Review* **40**(3): 276–299.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A. and Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning, *Decision Support Systems* **85**: 74–83.

Silva Filho, O. C., Ziegelmann, F. A. and Dueker, M. J. (2014). Assessing dependence between financial market indexes using conditional time-varying copulas: Applications to value at risk (var), *Quantitative Finance* **14**(12): 2155–2170.

Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T. M. (2018). Big data: Deep learning for financial sentiment analysis, *Journal of Big Data* **5**(1): 1–25.

Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, prediction, and search*, MIT press.

Tsai, M.-F. and Wang, C.-J. (2017). On the risk prediction and analysis of soft information in finance reports, *European Journal of Operational Research* **257**(1): 243–250.

Walker, C. B. (2016). The direction of media influence: Real-estate news and the stock market, *Journal of Behavioral and Experimental Finance* **10**: 20–31.

Wang, J. (2020). An intuitive tutorial to gaussian processes regression, *arXiv preprint arXiv:2009.10862* .

Wankhade, M., Rao, A. C. S. and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges, *Artificial Intelligence Review* **55**(7): 5731–5780.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, Vol. 2, MIT press Cambridge, MA.

Wilms, I., Gelper, S. and Croux, C. (2016). The predictive power of the business and bank sentiment of firms: A high-dimensional granger causality approach, *European Journal of Operational Research* **254**(1): 138–147.

Wu, B., Wang, L., Wang, S. and Zeng, Y.-R. (2021). Forecasting the u.s. oil markets based on social media information during the covid-19 pandemic, *Energy* **226**: 120403.
URL: https://www.sciencedirect.com/science/article/pii/S0360544221006526

A. Additional tables

This appendix has some additional tables with information about the number of articles segmented by topic, Table A1, and type of sentiment, Table A2. Moreover, there is also the Table A3 with the results of the Granger test, as discussed in Section 4.3.

Table A1
Issue per year

Year	Commodities	Currencies	Economic Indicators	Economy	Stock Markets
2015	1621	1612	696	0	4818
2016	3576	2762	1955	0	11369
2017	3578	3025	2308	0	14128
2018	3609	2122	2088	5678	17674
2019	3739	1916	2464	8404	17983
2020	3751	2126	2622	8269	17420
2021	3407	906	2126	8997	23469
2022	4186	903	2760	13720	28607
2023	565	146	544	2321	5169
Total	28032	15518	17563	47389	140637

Table A2
Sentiment per year

Year	Negative	Neutral	Positive
2015	5908	974	1865
2016	12311	2960	4391
2017	12953	4452	5634
2018	18011	6552	6608
2019	19847	7916	6743
2020	19673	7537	6978
2021	17600	10169	11136
2022	27090	10747	12339
2023	4646	1818	2281
Total	138039	53125	57975

Table A3
Granger test results

Direction	Lag	P-Value
S&P 500 index → Stock sentiment	**2**	***** 1.54e-11**
S&P 500 index → Stock sentiment	3	*** 1.59e-18
S&P 500 index → Stock sentiment	4	*** 8.99e-23
S&P 500 index → Stock sentiment	5	*** 2.86e-20
DXY → Currencies sentiment	2	** 0.0133
DXY → Stock sentiment	3	** 0.0202
DXY → Currencies sentiment	3	* 0.0544
DXY → Stock sentiment	4	* 0.083
DXY → Commodities sentiment	5	* 0.0824
DXY → Stock sentiment	5	** 0.0407
Brent crude oil → Commodities sentiment	**1**	***** 0.000283**
Brent crude oil → Stock sentiment	**1**	***** 0.000173**
Brent crude oil → Commodities sentiment	2	*** 0.000353
Brent crude oil → Stock sentiment	2	*** 2.78e-05
Brent crude oil → Commodities sentiment	3	*** 0.00277
Brent crude oil → Stock sentiment	3	*** 0.00302
Brent crude oil → Commodities sentiment	4	** 0.0421
Brent crude oil → Stock sentiment	4	*** 0.00235
Brent crude oil → Stock sentiment	5	** 0.0289
Stock sentiment → DXY	1	** 0.0339
Stock sentiment → Brent crude oil	2	* 0.0944
Commodities sentiment → Brent crude oil	2	*** 0.00362
Commodities sentiment → Brent crude oil	3	** 0.0127
Commodities sentiment → Brent crude oil	4	** 0.0243
Commodities sentiment → Brent crude oil	5	*** 0.00977
Economic indicators sentiment → Brent crude oil	1	* 0.0978
Economy sentiment → Brent crude oil	2	*** 0.00714
Economy sentiment → Brent crude oil	**3**	**** 0.0266**
Economy sentiment → Brent crude oil	4	** 0.033
Economy sentiment → Brent crude oil	5	** 0.0365

Notes: The lags are in day units. Significance levels < 0.01 ***, < 0.05 **, < 0.1 *. Results in bold indicate the same results found in the LPCMCI model.