ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Costola, Michele; Nofer, Michael; Hinz, Oliver; Pelizzon, Loriana

Working Paper

Machine learning sentiment analysis, Covid-19 news and stock market reactions

SAFE Working Paper, No. 288

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Costola, Michele; Nofer, Michael; Hinz, Oliver; Pelizzon, Loriana (2020) : Machine learning sentiment analysis, Covid-19 news and stock market reactions, SAFE Working Paper, No. 288, Leibniz Institute for Financial Research SAFE, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/224131

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Michele Costola | Michael Nofer | Oliver Hinz | Loriana Pelizzon

Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions

SAFE Working Paper No. 288

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Electronic copy available at: https://ssrn.com/abstract=3690922

Machine learning sentiment analysis, COVID-19 news and stock market reactions*

Michele Costola†
Michael Nofer‡
Oliver Hinz§
Loriana Pelizzon¶

September 11, 2020

Abstract

The possibility to investigate the impact of news on stock prices has observed a strong evolution thanks to the recent use of natural language processing (NLP) in finance and economics. In this paper, we investigate COVID-19 news, elaborated with the "Natural Language Toolkit" that uses machine learning models to extract the news' sentiment. We consider the period from January till June 2020 and analyze 203,886 online articles that deal with the pandemic and that were published on three platforms: MarketWatch.com, Reuters.com and NYtimes.com. Our findings show that there is a significant and positive relationship between sentiment score and market returns. This result indicates that an increase (decrease) in the sentiment score implies a rise in positive (negative) news and corresponds to positive (negative) market returns. We also find that the variance of the sentiments and the volume of the news sources for Reuters and MarketWatch, respectively, are negatively associated to market returns indicating that an increase of the uncertainty of the sentiment and an increase in the arrival of news have an adverse impact on the stock market.

Keywords: COVID-19 news, Sentiment Analysis, Stock Markets

JEL Classification: G10, G14, G15

*Loriana Pelizzon gratefully acknowledge research and financial support from the Leibniz Institute for Financial Research SAFE, Frankfurt. The usual disclaimer applies.

†Ca' Foscari University of Venice, michele.costola@unive.it.

‡Goethe University Frankfurt, michael@nofer.net.

§Goethe University Frankfurt, hinz@wiwi.uni-frankfurt.de.

¶SAFE and Ca' Foscari University of Venice, pelizzon@safe.uni-frankfurt.de.

Electronic copy available at: https://ssrn.com/abstract=3690922

1 Introduction

Non-financial news might have an impact on the stock market either because they provide information about fundamental asset values or because they are simply linked to investor psychology and sociology and therefore have a short term impact on stock returns, in line with theoretical models of noise and liquidity traders. The possibility to investigate the impact of news on stock prices has observed a strong evolution thanks to the recent use of natural language processing (NLP) in finance and economics. In time of market distress, the reaction of stock prices to news is even more pronounced and the recent COVID-19 pandemic of 2020 is the first global financial turmoil that could be analysed with NLP and extract the sentiment of investors regarding a specific topic and analyze the corresponding stock market reaction. In this paper, we attempt to investigate COVID-19 news, elaborated with Natural Language Toolkit, a suite of libraries and programs for symbolic and statistical natural language processing (NLP), that allows to use machine learning models to extract news' sentiment. We consider the period from January till June 2020 and analyze 203,886 online articles that deal with COVID-19 and that were published on three platforms: MarketWatch.com, Reuters.com and NYtimes.com. The algorithm we use allows us (i) to investigate how these three sources of information report news about COVID-19 and (ii) classify them as positive, negative or neutral using machine-learning based NLP methods and test whether these news have a positive, a negative or a neutral impact on the stock market. Given that the three news sources are based in the United States, we investigate if we observe any significant relationship of the SP500 returns after news on COVID-19 has been published on the platforms we considered. In this respect, we analyse three components for each sentiment: (i) the sentiment score, that represent the difference between positive and negative news sentiment on a particular day, (ii) the variance of this measure and (iii) the number of news on that day related to COVID-19.

Our findings show that there is a significant and positive relationship between the sentiment score and the market returns. This indicates that an increase (decrease) in the sentiment score implies a rise in positive (negative) news and corresponds to

Electronic copy available at: https://ssrn.com/abstract=3690922

positive (negative) market returns. We also find that the variance of the sentiments and the volume of the news sources for Reuters and MarketWatch, respectively, are negatively associated to market returns indicating that an increase of the uncertainty of the sentiment and an increase in the arrival of news have an adverse impact on the stock market. The results hold after controlling for a set of control variables such as the volatility index (VIX) which is a forward-looking estimate of future stock market volatility, the OFR Financial Stress Index as a measure of the stress in the global market, the growth rate for the COVID-19 cases which measures the pandemic trend, and the worldwide Google web-searches for the coronavirus topic as a proxy for the public attention.

This paper is related to the recent growing literature that investigates the impact of COVID-19 on the stock market and in particular to Mamaysky (2020) that documents the hypersensitivity of stock markets to news regarding the COVID-19 pandemic. Baker et al. (2020) that investigate the stock market reaction to previous pandemics, like the Spanish flu and SARS, with respect to the COVID-19 pandemic. Costola et al. (2020) and Smales (2020) that study the relationship between the google search volume and stock market returns. Gormsen and Koijen (2020) investigate how equity dividend futures and stock market levels allow to extract investor expectation about economic growth and Hartley and Rebucci (2020), analyze the market reaction to monetary policy interventions due to the COVID-19 pandemic crisis.

The paper is organized as follows. Section 2 describes the COVID-19 news sentiment construction through the data collection and the implemented methodology. Section 3 investigates the relationship between the extracted COVID-19 news sentiments and the financial markets. Finally, Section 4 concludes the paper.

2 COVID-19 News Sentiment construction

In the following section, we present the data collection procedure and the related source and the implemented methodology to extract Covid-19 news sentiment indicators.

Electronic copy available at: https://ssrn.com/abstract=3690922

2.1 Data Collection

Overall, we have collected 203,886 online articles that were published on three platforms between January 23, 2020 and June 22, 2020. Table 1 depicts summary statistics on the collected articles. Reuters.com and nytimes.com are the websites of the respective international news companies owned by Thomson Reuters and New York Times Company. The covered topics include business, politics, financial markets, science or health. In addition, we have also collected data from MarketWatch, which purely focuses on financial news and stock market data. The MarketWatch articles contain the most words on average (706) and the lowest maximum count (3857).

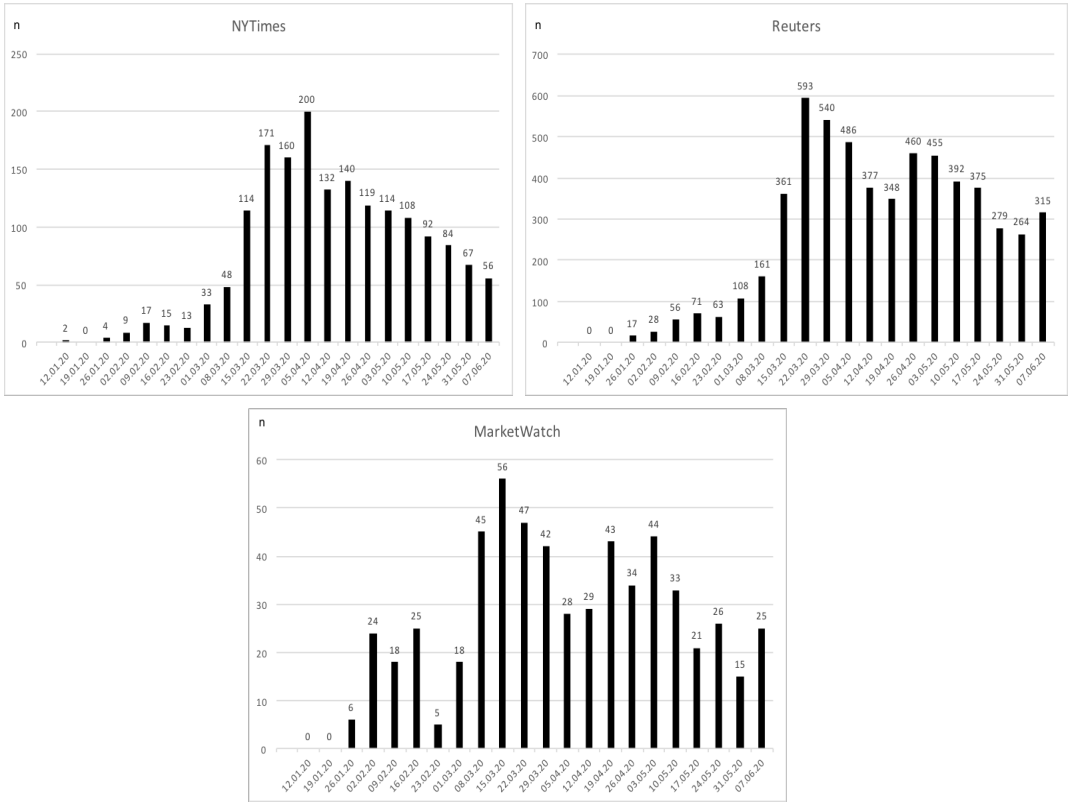
Table 1: Summary statistics

Website	Overall articles	Covid-19 related Articles	Average number of words	Maximum number of words	Method
Reuters.com	95378	4801	461	4607	Crawling
NYTimes.com	43172	1710	381	5859	API
MarketWatch.com	65336	588	706	3857	Crawling

The data collection process consists of three steps. First, we gather the URLs of the online articles either through the API or web crawling. The Reuters and MarketWatch crawler are developed using a link extractor written in Python Scrapy. The main goal of web scraping is to extract structured data from unstructured web pages. Scrapy contains the Spider class which can be used to define how to crawl and parse pages to extract items from a particular site (e.g., specifying the links). In addition, the Item class supports the creation of a container to collect the scraped data. The API and the crawler allow us to store the meta data in the database, such as headline, author, publish date and URL. Afterwards, we filter the COVID-19 URLs by focusing on the related keywords, such as 'COVID" and "Corona". In the last step, we collect all text elements (p-tags) from the remaining URLs, i.e. date, title, author and text. Figure 1 depicts the weekly number of collected articles on NYTimes, Reuters and MarketWatch during the course of the pandemic.

Electronic copy available at: https://ssrn.com/abstract=3690922

Figure 1: Weekly number of collected articles over time for the three news sources (MarketWatch, Reuters and NYTimes)



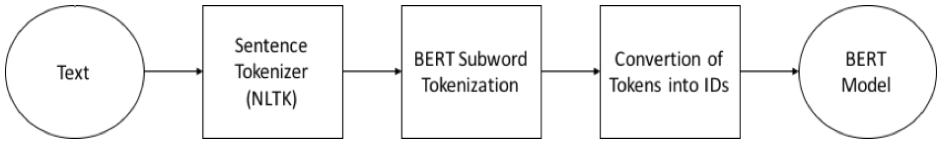
2.2 News Sentiment Analysis

After the data collection, we tokenize all articles using the Natural Language Toolkit (NLTK), one of the leading platforms for building Python programs to work with human language data. Figure 2 displays the preprocessing. In general, tokenization means dividing a big quantity of text into smaller parts called tokens. Machine learning models need numeric data to be trained and make a prediction. Word tokenization becomes a crucial part of the text (string) to numeric data conversion. Following the tokenization, we can compute the "compounded sentiment score" using the BERT-based Financial Sentiment Index (Hiew et al., 2019). This textual-based sentiment index relies on BERT (Bidirectional Encoder Representations from Transformers) originally developed by Google and Devlin et al. (2018).

BERT is an open-source model that was pre-trained with millions of words from the entire Wikipedia corpus, employing a bidirectional Transformer encoder to predict masked words. The model performs two tasks. First, BERT randomly masks a fraction of

Electronic copy available at: https://ssrn.com/abstract=3690922

Figure 2: The preprocessing steps using the Natural Language Toolkit (NLTK)



words and predicts the words that have been masked-out. The "next sentence prediction" aims to predict whether the other sample of words follows the first one (Devlin et al., 2018).

Due to this pre-training approach, BERT has outperformed various state-of-the art NLP techniques. For instance, it achieves an accuracy of 94.9 percent in the completion of the Stanford Sentiment Treebank. Hiew et al. (2019) apply BERT for the stock market prediction and we use this pre-trained model for the sentiment analysis.

All tokenized sentences are classified as positive, negative or neutral along with the respective probabilities (logits). The sentiment is calculated as follows:

$$Sentiment = Logit_{Positive} - Logit_{negative} \tag{1}$$

The Sentiment is therefore the probability that the sentence is positive minus the probability that the sentence is negative.

The following text excerpt of a collected article illustrates the calculation of the sentiment:

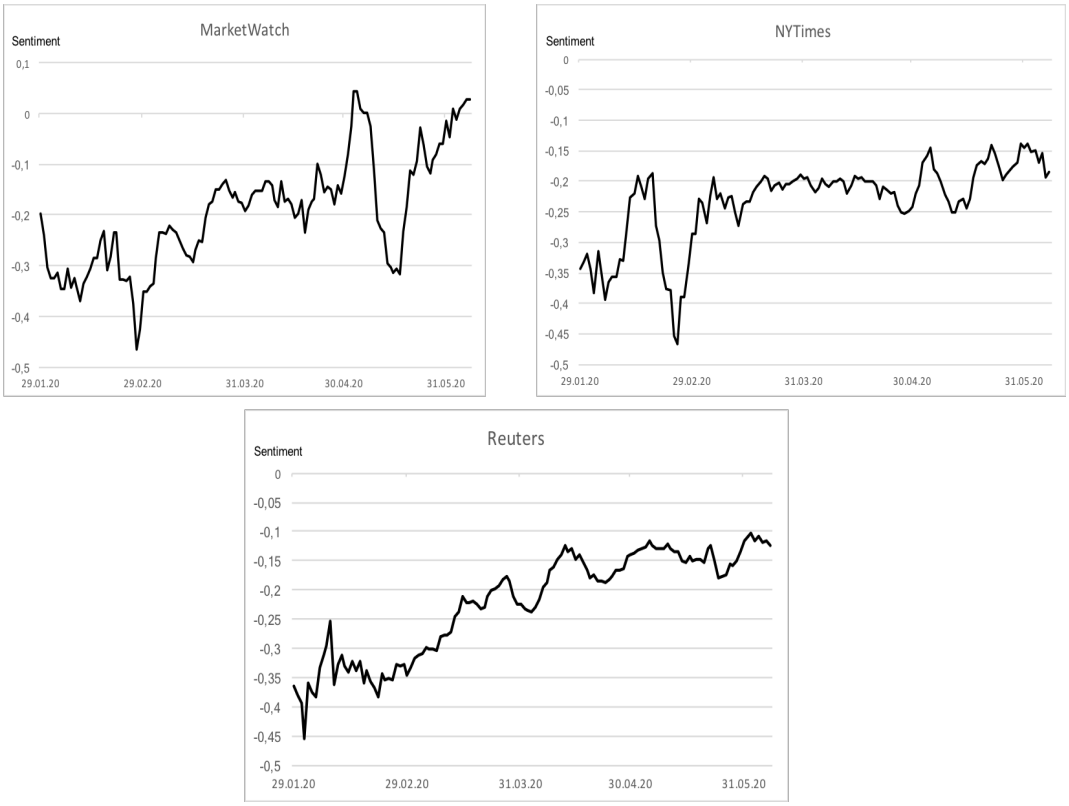
> "The latest tranche of $60 million was raised a month ago from existing investors who wanted to provide more capital as they saw the business booming with the pandemic, said Gustavo Sapoznik, founder and chief executive of ASAPP."

The BERT model classifies this message into a positive value of 0.843, a negative value of 0.007841 and a neutral value of 0.1484. This results in a sentiment of 0.835 which represents a positive prediction of the model. The sentiment for an entire article is determined by averaging the individual sentiment of each sentence contained therein. Figure 3 depicts the rolling 7-day average of the sentiment over time. The sentiment is

Electronic copy available at: https://ssrn.com/abstract=3690922

mostly negative across all platforms, however we observe an upward trend during the course of the pandemic. There's a sharp decline of the sentiment in late February when the western world started to realize that COVID-19 could be a global threat to the economy.

Figure 3: The sentiment indicators over time according the three news sources (MarketWatch, NYTimes and Reuters)



If we compare the weekly correlations of sentiments between the platforms as shown in Table 2, it turns out that the Reuters and NYTimes sentiment is correlated most. This might be explained by the nature of the websites. In contrast to MarketWatch, which focuses on financial news, NYTimes and Reuters cover a broader range of topics (e.g., economics, science, etc.).

Table 2: Weekly correlations of sentiment scores

	MarketWatch	NYTimes	Reuters
MarketWatch	1.000	0.081	0.338
NYTimes	0.081	1.000	0.876
Reuters	0.338	0.876	1.000

Besides the sentiment, we will include the daily variance of the sentiment and the

Electronic copy available at: https://ssrn.com/abstract=3690922

volume of articles as additional measures for COVID-19 news sentiment. The volume simply represents the number of COVID-19 related news articles on day t. The variance of the sentiment is the average of the squared differences from the mean value on day t.

3 Stock market and the COVID-19 news sentiment

In this section, we investigate the relationship between the extracted COVID-19 news sentiments and the financial markets. Given that the three considered news platforms are predominantly based in US, we focus on this financial market and consider the returns on the S&P 500. The analysed period is between January 23, 2020 to June 22, 2020 at daily frequency. In this respect, we consider for each COVID-19 news sentiment three daily components: i) the sentiment (Levels); ii) the daily variance of the sentiment (σ^2); and iii) the volume of the articles for the related news source (Volume). Furthermore, we add a set of control variables that could exert an impact on the stock market returns: i) the volatility index (VIX); ii) the OFR Financial Stress Index[1] which measures the stress in global financial markets; iii) the growth rate of the COVID-19 world cases (grCOVID-19);[2] and iv) The worldwide web searches on Google matching the "*coronavirus*" topic (Google Trends) to proxy the general public attention as in Costola et al. (2020).

Table 3 includes the estimates for the returns on the S&P 500. The estimates in models 1-3 include the daily market sentiment, the variance of the sentiment, the volume of articles for each news source (MarketWatch, NYTimes and Reuters) and the set of control variables previously defined. All the sentiment indicators are significant and positive related to the returns on S&P500 indicating that an increase (decrease) in the sentiment implies a rise in positive (negative) news and corresponds to positive (negative) market returns. This highlights that there is a relationship between the sentiment on COVID-19 and the stock market after controlling for the VIX (significant and negatively related), the OFR Financial Stress Index (significant and positively related only for MarketWatch),

[1]The index is proposed by the Office Financial Research and is available at https://www.financialresearch.gov/financial-stress-index/.

[2]Data has been downloaded by Bloomberg using the mnemonic NCOVCNCA.

Electronic copy available at: https://ssrn.com/abstract=3690922

the growth rate for the COVID-19 cases (not significant for all) and the worldwide Google web-searches for coronavirus (significant and positively related for all). Regarding the variance of the sentiments and the volume of the news sources, the only significant variable is the volume for MarketWatch which is negatively associated to market returns indicating that an increase in the arrival of news is mostly related to a negative market sentiment during the considered period. If we consider the three sentiments together (Model 4), results confirm the previous findings for MarketWatch and NYTimes except for Reuters which is no longer significant. Most likely, the correlation between Reuters and NYTimes leads to the result that the variable for NYTimes catches this effect alone. Finally, we include in model 5 all the sentiments with their associated variances and volumes. The results for the sentiments remain unchanged with respect to model 4. Regarding the variance of the daily sentiments, we find that Reuters is significant and negatively related to market returns. This behavior is similar to the leverage effect between returns and volatility as in asset pricing. In the current case, we read that an increase of uncertainty in the sentiment negatively affects the stock market returns. The results for the specified control variables remain substantially unchanged.

Additionally, we have repeated the analysis by using the principal component analysis (PCA) on the three news sources by including in the regression the first principal component on the sentiment scores, variances and article volumes. Results are included in Appendix A and confirm the significant and positive relationship between the (first component on the) sentiment scores and market returns.

4 Conclusion

This paper investigates the relation between the stock market and news sentiment related to COVID-19. In this respect, we consider the Natural Language Toolkit that uses machine learning models to extract COVID-19-related news sentiment on 203,886 online articles that were published on MarketWatch, Reuters and NYtimes between January and June 2020. Our findings show that there exists a significant and positive relationship

Electronic copy available at: https://ssrn.com/abstract=3690922

Table 3: Estimates on the S&P 500 returns using the COVID-19 news sentiments (MarketWatch, NYTimes and Reuters) plus a set of control variables (the VIX, the OFR index, the growth rate on COVID-19 world cases and the web-searches on Google for the coronavirus topic). Notes: HAC standard errors are in parentheses; *, **, and *** denote statistical significance at 10%, 5%, and 1%, respectively.

	1	2	3	4	5
const	0.0736***	0.0747***	0.0736***	0.0807***	0.0843***
	(0.0183)	(0.0183)	(0.0218)	(0.0198)	(0.0182)
MarketWatch	0.0373***			0.0387***	0.0385***
	(0.0108)			(0.0112)	(0.0113)
MarketWatch(σ^2)	-0.0273				0.0091
	(0.0566)				(0.0603)
MarketWatch(Volume)	-0.0018*				-0.0016
	(0.0011)				(0.001)
NYTimes		0.0599***		0.0434**	0.047**
		(0.0223)		(0.0165)	(0.0184)
NYTimes(σ^2)		0.0247			0.0403
		(0.0563)			(0.0759)
NYTimes(Volume)		0.0000			0.0000
		(0.0009)			(0.0009)
Reuters			0.0668***	0.042	0.0288
			(0.0233)	(0.0278)	(0.027)
Reuters(σ^2)			-0.0943		-0.1017*
			(0.066)		(0.0567)
Reuters(Volume)			0.0001		0.0002
			(0.0001)		(0.0001)
VIX	-0.0029***	-0.0031***	-0.0029***	-0.0031***	-0.0032***
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0007)
OFR	0.0045***	0.0039	0.0016	0.0017	0.0013
	(0.0016)	(0.0032)	(0.0018)	(0.0016)	(0.003)
grCOVID-19	-0.0091	-0.0051	-0.0124	0.0008	-0.0009
	(0.0163)	(0.0106)	(0.0157)	(0.0127)	(0.0132)
Google Trends	0.0011**	0.0011**	0.0013***	0.0015***	0.0016***
	(0.0005)	(0.0004)	(0.0004)	(0.0005)	(0.0004)
Obs	105	105	105	105	105
adj-R2	0.2273	0.208	0.198	0.2556	0.2611
F-stat	5.3701	4.9019	4.6671	6.1006	3.8266
p-value	3.1964e-05	9.0607e-05	0.00015346	6.4611e-06	6.6351e-05

Electronic copy available at: https://ssrn.com/abstract=3690922

between sentiment score and the market returns. This highlights that the flow of news on COVID-19 has had an impact on the financial market since they contributed to form market participants' expectations about the evolution of the pandemic, the real economy and the stock market.

Electronic copy available at: https://ssrn.com/abstract=3690922

References

Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., Viratyosin, T., 2020. The unprecedented stock market reaction to COVID-19. The Review of Asset Pricing Studies .

Costola, M., Iacopini, M., Santagiustina, C., 2020. Public Concern and the Financial Markets during the COVID-19 outbreak. Available at SSRN 3591193 .

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

Gormsen, N. J., Koijen, R. S., 2020. Coronavirus: Impact on stock prices and growth expectations. University of Chicago, Becker Friedman Institute for Economics Working Paper .

Hartley, J. S., Rebucci, A., 2020. An Event Study of COVID-19 Central Bank Quantitative Easing in Advanced and Emerging Economies. Tech. rep., National Bureau of Economic Research.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., Xu, Y., 2019. BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. arXiv preprint arXiv:1906.09024 .

Mamaysky, H., 2020. Financial Markets and News about the Coronavirus. Available at SSRN 3565597 .

Smales, L. A., 2020. Investor Attention and Global Market Returns during the COVID-19 Crisis. Available at SSRN 3623915 .

Electronic copy available at: https://ssrn.com/abstract=3690922

A Stock market and the PCA on COVID-19 news sentiment

In this section, we perform the analysis by using PCA on the three news sources and consider the first principal component on the sentiment scores, variances and article volumes. Results confirms the first component on the sentiment scores is significant and positive related to market returns.

Table 4: Estimates on the S&P 500 returns using PCA on the COVID-19 news sentiments (MarketWatch, NYTimes and Reuters) plus a set of control variables (the VIX, the OFR index, the growth rate on COVID-19 world cases and the web-searches on Google for the coronavirus topic). Notes: HAC standard errors are in parentheses; *, **, and *** denote statistical significance at 10%, 5%, and 1%, respectively.

const	0.0669***
	(0.0174)
Score-PCA1	0.0086***
	(0.003)
σ^2-PCA1	-0.0023
	(0.0033)
Volume-PCA1	-0.0039
	(0.0052)
VIX	-0.0031***
	(0.0009)
OFR	0.0069***
	(0.0026)
grCOVID-19	-0.014
	(0.0157)
Google Trends	0.0009**
	(0.0005)
Obs	105
adj-R2	0.2152
F-stat	5.0737
p-value	6.1719e-05

Electronic copy available at: https://ssrn.com/abstract=3690922

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 287	Kevin Bauer, Nicolas Pfeuffer, Benjamin M. Abdel-Karim, Oliver Hinz, Michael Kosfeld	The Terminator of Social Welfare? The Economic Consequences of Algorithmic Discrimination
No. 286	Andreass Hackethal, Michael Kirchler, Christine Laudenbach, Michael Razen, Annika Weber	On the (Ir)Relevance of Monetary Incentives in Risk Preference Elicitation Experiments
No. 285	Elena Carletti, Tommaso Oliviero, Marco Pagano, Loriana Pelizzon, Marti G. Subrahmanyam	The COVID-19 Shock and Equity Shortfall: Firm-Level Evidence from Italy
No. 284	Monica Billio, Michele Costola, Iva Hristova, Carmelo Latino, Loriana Pelizzon	Inside the ESG Ratings: (Dis)agreement and Performance
No. 283	Jannis Bischof, Christian Laux, Christian Leuz	Accounting for Financial Stability: Bank Disclosure and Loss Recognition in the Financial Crisis
No. 282	Daniel Munevar, Grygoriy Pustovit	Back to the Future: A Sovereign Debt Standstill Mechanism IMF Article VIII, Section 2 (b)
No. 281	Kevin Bauer	How did we do? The Impact of Relative Performance Feedback on Intergroup Hostilities
No. 280	Konstantin Bräuer, Andreas Hackethal, Tobin Hanspal	Consuming Dividends
No. 279	Tobin Hanspal, Annika Weber, Johannes Wohlfart	Exposure to the COVID-19 Stock Market Crash and its Effect on Household Expectations
No. 278	Sandra Eckert	EU Agencies in Banking and Energy Between Institutional and Policy Centralisation
No. 277	Dominique M. Lammer, Tobin Hanspal, Andreas Hackethal	Who Are the Bitcoin Investors? Evidence from Indirect Cryptocurrency Investments
No. 276	Massimiliano Caporin, Loriana Pelizzon, Alberto Plazzi	Does Monetary Policy Impact International Market Co-Movements?
No. 275	Loriana Pelizzon, Max Riedel, Zorka Simon, Marti Subrahmanyam	Collateral Eligibility of Corporate Debt in the Eurosystem

Electronic copy available at: https://ssrn.com/abstract=3690922