The CoRisk-Index: Measuring economic risks related to COVID-19 in real time

Fabian Braesemann (✉ fabian.braesemann@sbs.ox.ac.uk)

University of Oxford <https://orcid.org/0000-0002-7671-1920>

Fabian Stephany

University of Oxford

Leonie Neuhäuser

Hertie School of Governance

Niklas Stoehr

ETH Zurich

Philipp Darius

Hertie School of Governance

Ole Teutloff

Hertie School of Governance

Article

Keywords: COVID 19, economic risk, CoRisk-Index

Posted Date: August 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-81992/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Humanities & Social Sciences Communications on February 2nd, 2022. See the published version at <https://doi.org/10.1057/s41599-022-01039-1>.

The *CoRisk*-Index: Measuring economic risks related to COVID-19 in real time

ABSTRACT

The global spread of Covid-19 has caused major economic disruptions. Governments around the world provide considerable financial support to mitigate the economic downturn. However, effective policy responses require reliable data on the economic consequences of the corona pandemic. We propose the *CoRisk*-Index: a real-time economic indicator of Covid-19 related risk assessments by industry. Using data-mining and natural language processing techniques, we analyse over 5,800 reports from US companies representing more than 44 million US employees. We construct three measures: (a) the number of 'corona' words in each report, (b) the share of corona-mentioning firms per industry, and (c) the average *text negativity* of the sentences mentioning corona in each report, which are aggregated in the *CoRisk*-Index. The index correlates with U.S. unemployment data and preempts stock market losses of February 2020. Moreover, the *CoRisk*-Index can be generalised to retrace past shocks such as the U.S. - China trade war in 2018. The data presented here help researchers and policy-makers to better understand industry-specific economic risks related to Covid-19 and ultimately to better forecast and manage the unfolding economic crisis. For immediate access to the data, we provide all findings and raw data on an interactive online dashboard in real time.

Introduction

The Covid-19 pandemic has caused the largest global economic disruption of the 21st century¹⁻³. Travel bans, supply chain failures, and store closures pose significant risks to entire industries. In an attempt to mitigate the general economic downturn, governments are mobilising considerable financial support^{4,5}. While immediate general support programmes can help to stabilise economies ad-hoc, targeted and continuous countermeasures require reliable and up-to-date information on economic trends. However, the fastly changing dynamics of the Covid-19 pandemic pose a challenge for economic forecasting and targeted policy intervention.

In the absence of timely empirical data, investigations on the economic impact of Covid-19 explore two pathways. One stream of research relies on metrics based on past economic shocks, such as economic simulations^{6,7}. Another stream, following the paradigm of computational social science⁸, explores alternative data sources, such as stock markets⁹⁻¹¹, news articles¹², website content¹³ or search queries¹⁴.

However, both approaches come with limitations in the current situation. Traditional economic indicators, like composite indices¹⁵ or labour market statistics have a sizeable time lag. Empirical data with a high time resolution, such as stock markets, search queries, and news articles include a lot of noise and are prone to herd behaviour. Moreover, most metrics do not allow for a granular analysis of the economic repercussions by industry. Examining current stock market dynamics, for example, reveals a general economic downturn, but it does not allow to isolate the sector and domain-specific risks related to the economic threats of the pandemic. First attempts to quantify risk perceptions related to Covid-19 on the firm-level are limited to the company perspective¹⁶ and lack the relation to the macro-economic environment.

In response, we propose a real-time analysis of business reports (10-K reports) filed to the U. S. Securities and Exchange Commission (SEC) using data-mining and natural language processing techniques. SEC filings represent financial statements of publicly listed companies including a risk assessment. These reports are imperative to comply with legal and insurance requirements and therefore contain reliable risk evaluations. Prior work has underlined the forward-looking nature of SEC filings, as they allow for an effective prediction of volatility on stock market returns and other financial metrics^{17–19}.

In collecting all 10-K reports published from 30th January 2020 — the day the term *coronavirus* first appeared in a 10-K report — to 30th June, we assess and track the reported risk assessments related to Covid-19 for the different sectors¹ of the US economy. The 5,800 reports analysed here cover the risk assessments of U.S. companies representing more than 44 million U.S. employees, i. e. one-third of the total U.S. economy. We construct three measures: (a) the number of 'corona' words in each 10-K report, (b) the share of corona-mentioning firms per industry, and (c) the average *text negativity* of the sentences mentioning corona in each report. The text negativity is calculated as the share of negative words²⁰ per corona-mentioning sentence. The geometric mean of corona-words and text negativity builds the CoRisk-Index, which is calculated for each industry. Moreover, to understand how companies in different industries assess Covid-19 as a risk over time, we use natural language processing to identify the topical context in which Covid-19 related risks are described.

As we will demonstrate, the CoRisk-Index represents a highly informative, reliable, up-to-date and low-cost data source. The data can help researchers and policy-makers to better understand and forecast the economic repercussions related to COVID-19 in different industries and to design better policy responses. For immediate access to all raw data we offer an interactive online dashboard.²

The remaining paper proceeds as follows. In the next section, we introduce the data and CoRisk-Index in more detail. We show that the CoRisk measures relate to stock market and unemployment trends. Moreover, we demonstrate that the CoRisk-Index can be generalised to retrace past shocks such as the U.S. - China trade war in 2018. Subsequently, we discuss policy implications and critically reflect upon the limitations of our approach. The Methods section provides a comprehensive step-by-step explanation of the data collection, pre-processing and analysis.

Results

Anticipation of the economic shock

The business risk measures captured by the CoRisk-Index anticipated the economic reactions of the unfolding crisis before the number of Covid-19 cases surged in the United States (Fig. 1A). Before stock markets plummeted on February 20th (Fig. 1B), the share of 10-K reports mentioning corona has been rising steadily (Fig. 1C). In early May, when the number of new infections started to decrease, the share of corona-mentioning reports established itself on a high level of just below 80 %, before rising further in June, when the number of new cases started to rise again.

Similarly, the average share of negative words in sentences mentioning 'corona' or 'covid' (*text negativity*) spiked in early February (Fig. 1D), peaking at a share of 7 % negative words per corona-sentence just before the most severe stock market losses ('Black Monday') on March 9th. The text negativity measures the sentiment with which firms report about Covid-19 related business issues. The sharp increase reflects the uncertainty in global economic outlooks, which businesses suddenly became aware of after the pandemic turned from a regional to a global health emergency. In the subsequent months, the text negativity steadily decreased, supported by economic aid packages and the release of stay-at-home orders, declining to a level of around 5 % recently.

The CoRisk-Index

Based on the count of 'corona' keywords and the text negativity, we construct the CoRisk-Index (Fig. 2A) for eight out of 13 sectors of the U.S. economy, which are well represented by SEC filings (see Methods section). The index serves as a real-time economic indicator measuring industry risks related to Covid-19.

¹In the paper, we use the term sector and industry interchangeably; however, we want to emphasise that the level of analysis is two-digit sectors of the U.S. economy, according to the North American Industry Classification System. As all raw data are being made publicly available, the analysis could be reproduced on the level of individual firms or four-digit industries and sub-sectors.

²The data and interactive dashboard can be found at <http://oxford.berlin/corisk/>.

Figure 1. The CoRisk measures anticipate the expected economic shocks of the Covid-19 pandemic. (A) Daily cases of Covid-19 infections in the United States (source: www.ourworldindata.org): the cases started to rise sharply in March and April 2020. (B) S&P 1200 Global Index (source: www.spglobal.com): As the Covid-19 pandemic reaches the level of a global health crisis stock markets plummet on 20th February. (C) Share of companies mentioning 'corona' in 10-K reports: weeks before the pandemic hits the United States, the share of firms reporting about the virus starts to rise. (D) Share of negative words in 'corona' sentences of 10-K reports: The negative sentiment surges before the global stock market crash and peaks before "Black Monday's" most severe stock market losses.

Figure 2. The CoRisk-Index reveals the dynamics of the economic crisis in different industries.

(A) The CoRisk-Index: a composite measure of industry-specific business risks. It shows how industries differ in timing and magnitude of risk awareness towards Covid-19. The risk indices for Manufacturing and Wholesale & Retail climb early on, reaching levels higher than 125 index points by May; Finance's index grows more slowly and reaches a level of around 90 points by June. (B) CoRisk topics: the share of domain-specific keywords in 'corona' sentences of 10-K reports in two industries. The data reflects the relevance of the different dimensions of the crisis: Travel restrictions became a topic of reporting in Manufacturing in early February (top panel), before firms started to mention supply chain related issues in both industries in March (second panel). Demand disruptions caused by the lockdown have become a highly relevant issue for Retail during March and April (third panel); more recently, uncertainties regarding the governmental aid CARES act have been reported in both industries (lowest panel).

Sectors differ in their timing and magnitude of risk awareness. For example, Finance shows relatively low levels of risk awareness, while Manufacturing and Wholesale & Retail reveal early and strong risk sentiments. Overall, 75 % of the firms in manufacturing and 74 % of the retail companies have mentioned the coronavirus as a potential economic risk. Only 28 % of the businesses in finance have done so. The topical attribution of corona-mentioning sentences allows us to investigate how the pandemic influenced business perceptions related to the topics of travel restrictions, supply shocks, demand disruptions, and governmental aid (Fig.2B). As the U.S. government posed first travel restrictions on January 31st, travel-related reporting in Manufacturing surged followed by Wholesale & Retail a month later. Subsequently, both industries began to express significant concerns regarding supply delays throughout February and March. The supply disturbance was overshadowed by a substantial demand shock with the announcement of stay-at-home orders by mid-March. Since then, the demand disruptions particularly affected the reporting in the Wholesale & Retail sector. More recently, reporting in both industries started to deal with the implications of the "Coronavirus Aid, Relief and Economic Security (CARES) Act" (signed by President Trump on March 27th releasing more than USD 500 billions in loans to the U.S. economy).

Application & Generalisability

To validate the CoRisk data as a broadly applicable real-time economic indicator, we relate the weekly changes of the CoRisk-Index with the U.S. unemployment initial claims data (as a percentage of total

A An application of the CoRisk-Index: correlation with unemployment initial claims per industry

B Generalisation of the CoRisk methodology: share of negative words in 'china' sentences in 2018 SEC filings

Figure 3. The CoRisk data reflect macroeconomic developments. (A) The CoRisk-Index (weekly growth: $\text{CoRisk}_{t-1} / \text{CoRisk}_{t-2}$) and weekly unemployment initial claims in different industries (28th March to 09th May): changes of the CoRisk-Index are correlated with unemployment claims in 'All Industries', Manufacturing, and Wholesale & Retail. By May, the economic crisis has not led to a rise of unemployment in Finance. (B) Generalisability of the CoRisk methodology: share of negative words in 'china' sentences in 10-K reports during the 'US-China trade war' in 2018 (solid line and grey dotted trend line). The methodology allows to extract the china-related negative sentiment from the reports and it reflects negative (red) and positive (green) events during the crisis, pointing towards the general applicability of the tool to reveal granular insights about the effects of events on businesses.

unemployment per sector; see Methods section). Despite limited data availability (currently, in the very early phase of the crisis, not much unemployment data are available), Figure 3A reveals correlations between the two stationary time series for a number of industries. However, the crisis has not yet reached all sectors of the economy. Hence, the financial industry, for example, shows little reaction in terms of rising unemployment initial claims, which results in a very low correlation.

The CoRisk methodology can be generalised beyond the case presented here. Reports filed to the U.S. Securities and Exchange Commission represent a rich source of textual data reflecting political and economic events, which can be used to track and predict the influence of external events on the business environment. We exemplify the generalisability of the approach on the case of the U.S.-China trade war in 2018. Figure 3B shows the text negativity of 'china' sentences, i. e. share of negative words in sentences mentioning the keyword 'china'. The negativity varies between 1.7 % and 3.4 % over the course of the trade war, with tariff announcements being reflected by periods of sharply increasing negativity and relaxing events such as trade talks being reflected by decreasing text negativity. The example illustrates that the CoRisk methodology provides a flexible tool that is able to track business risk factors in numerous applications. Besides this example, we provide a comparison of the overall report negativity (share of negative words in 10-K reports) and macroeconomic variables in the appendix.

Discussion

Governments are eager to counterbalance the looming global economic crisis induced by the Covid-19 pandemic with cyclical and fiscal policy packages of enormous volumes. The CoRisk-Index represents an attempt to contribute to the pressing demand for empirical data on the economic impact of the ongoing pandemic. It equips policy-makers and researchers with a tool to analyse and track industry-specific risk perceptions related to Covid-19 in real time, based on data that represents a large part of the U.S. economy. Our analysis shows that industries vary greatly in their risk perception as well as with regards to the topics that they are concerned with. It underlines the necessity to incorporate alternative data sources of up-to-date empirical information in the current situation, in order to better understand the differential effects of Covid-19 on different sectors of the economy. The CoRisk data will help to forecast the economic effects of the crisis and to design viable support instruments targeting the most affected industries.

However, the analysis comes with potential limitations. Firstly, we limit our investigation to the United States given the proven reliability of 10-K reports as a source of company risk reporting¹⁷. Secondly, the value of the approach relies on company self-reporting. As many of the implications of the corona crisis are still uncertain, our approach thus reflects a way to approximate potential implications on current estimations of company experts in the different sectors. Naturally, this does not include risks that are unforeseeable for companies themselves at a given point in time. Thirdly, the CoRisk-Index is currently limited to the United States, as it appears to be the only country that provides company risk reports in a highly standardised and freely accessible format. Nevertheless, we are working on implementing the index for other countries.

Nonetheless, we believe that our measures represent a valuable source of empirical information about the issues faced by different industries in the current situation. The rapidly changing global pandemic calls for a constant re-assessment of economic repercussions and the corresponding policy responses. In this situation, agile computational social science methods such as data-mining and natural language processing allow us to extract valuable information at low-cost and in real time. As the crisis unfolds, we continuously extend our work. The CoRisk-Index is updated weekly until the end of the pandemic, in order to support fellow researchers and policy makers with timely data on industry-specific risk perceptions. In addition to the findings presented in this paper, we provide detailed industry-specific insights and all raw data (including filter- and download options) via an interactive online dashboard.³

³The data and interactive dashboard can be found at <http://oxford.berlin/corisk/>.

Methods

Data collection

Automated extraction of 10-K reports

The U. S. Securities and Exchange Commission (SEC) stores all reporting in a central repository⁴. Here, users can access meta-level information, such as index files, e.g., lists of all reports issued in the second quarter of 2020 directly. Alternatively, individual filings and meta-level information can be retrieved via various statistical packages, e.g., "*edgar*" (R) or "*sec-edgar-downloader*" (python), or a freemium API⁵.

However, several limitations make the use of these ready-made devices impractical for the specific research purposes of this work. Downloading all relevant reports as .txt/.html files via the ready-made packages is possible in theory but requires a lot of time and hard disk space. In addition, the .txt file contains several unwanted css/html code patterns that make the identification of corona-sentences and the counting of words from them unreliable.

Alternatively, we initially refer to the *crawler.idx* index file⁶ in the SEC repository. The index file holds a full list of all reports issued in a given quarter. This document shows the meta-level *index.htm* page for each company. On this page, the most recent 10-K report is linked⁷. Unfortunately, the htm-version of each 10-K report has a cryptic file name that can not be anticipated or guessed with the knowledge of company, date or industry parameters. Hence the currently implemented scraper (python 2.7) pipeline of our project, first, fetches the list of recently listed 10-K reports from *crawler.idx*, secondly, constructs the meta-level *index.htm* for each company and finds the link to the most recent 10-K report. The algorithm then scrapes the report text before, lastly, identifying the sentences related to corona (examples of text element that contain the term 'coronavirus' are displayed in Figure A4). All sentences and their respective report properties are stored for later processing.

Collection of stock market data

The collection of stock market data follows the incentive to provide reference data for the company SEC filings. We obtained stock market data through the *Yahoo! Finance API* following two successive steps: First, we extracted the CIK (central index key) identifier of each company from the SEC filings to compile a comprehensive list of all companies, their CIK and stock ticker identifiers. Using the stock tickers, we then retrieved the historic closing stock values of each company per trading day between January 1st 2018 and May 15th 2020.

Collection of unemployment data

Due to the recent outbreak of the crisis, there are not many data points of monthly industry-specific unemployment rates available, which are provided by the Bureau of Labour Statistics. The crisis had first labour market repercussions only in March; until 5th June, there are only unemployment rates for March (published in April) and April (published in May) available. Therefore, we consider a different data set to approximate the labour market repercussions of the pandemic. We use weekly unemployment initial claims data provided by the Economic Policy Institute.⁸ The data covers the number of weekly initial claims per sector in nine weeks from 14th March to 9th May 2020 for 19 US states. Additionally, the data contains information about the total employment per industry and state.

Using this dataset, we can calculate the share of weekly unemployment initial claims per total employment (the total employment is assumed to be constant in the observation period) in the 19 US states.

Specification of corona keywords

Since this study examines the attention attributed to COVID-19 in the SEC filings, the discovery mechanism of relevant COVID-19 mentions is of central importance. To mitigate susceptibility to errors due to word splitting, stemming and other text preprocessing, we decided for the most simple approach based on the matching of regular expressions. We scanned the reports for the two relatively unambiguous terms "corona"

⁴<https://www.sec.gov/Archives/edgar/full-index/>

⁵<https://sec-api.io/>

⁶<https://www.sec.gov/Archives/edgar/full-index/2020/QTR2/crawler.idx>

⁷E.g., <https://www.sec.gov/Archives/edgar/data/1084869/0001437749-20-002005-index.htm>

⁸Source: Economic Policy Institute (2020) "Weekly UI initial claims by state and industry" (Data collected from various sources), https://economic.github.io/ui_state_detailed/. We are grateful to Andrew Van Dam how has pointed us to this data source.

and "covid", also accounting for "coronavirus" and "covid-19" without duplication. For this process, the entire text is set to lower case.

Topic detection

We apply unsupervised methods for topic detection from Natural Language Processing to further analyse the reports with regard to topic related risks. Different sectors are facing different challenges, therefore companies are reporting about different corona-related risks. We aim to capture these risk topics via a keyword search on predefined topics. In order to explore possible topics, we used Latent Dirichlet Allocation (LDA) for unsupervised topic modelling, similar to Dyer et al. (2017)²¹. We only apply the topic model to corona-related paragraphs in the risk sections. We additionally examine the most frequent words and bi-grams in the documents. Using this exploratory analysis, we define a set of topics, which are specified by keywords. We then conduct a keyword search to count how much these terms are mentioned in the different industries in order to estimate the topic prevalence.

Unsupervised topic modelling

We use unsupervised learning techniques to explore the space of topics that companies discuss when describing coronavirus-related risks. Latent Dirichlet Allocation (LDA) is a Bayesian computational linguistic technique that identifies the latent topics in a corpus of documents²². This statistical model falls into the category of generative probabilistic modelling: a generative process which defines a joint probability distribution over the observed random variable, i.e. the words of the documents, and the hidden random variables, i.e. the topic structure. In other words, LDA uses the probability of words that co-occur within documents to identify sets of topics and their associated words²¹. The number of topics has to be defined in advance. LDA is a frequently used technique to identify main topics in a corpus. Nevertheless, the interpretation of these topics can sometimes be difficult. We thus perform LDA for explorative purposes in our research only and apply the following steps:

Sample restriction We filter all sentences from the risk sections that mention either "corona" and "covid", thereby also accounting for "coronavirus" and "covid-19".

Text preparation Before we train the LDA model we prepare the documents to achieve better performance of the method. We remove all common English stopwords, which are frequent words such as "is," "the," and "and" as well as those words which appear in at least 80% of the documents. These words are not useful in classifying topics as they are too frequent and therefore decrease performance. Moreover, we delete all words that do not occur in at least two documents.

LDA We turn the documents into numerical "Bag of words" feature vectors, disregarding word order. We then use LDA to extract the topic structure. Like any unsupervised topic model, this requires setting the number of topics a priori. We selected this key parameter based on semantic coherence, evaluating a range of two to eight topics leading to a final model of four topics. The top ten terms of each topic are displayed in Table A1.

Dictionary-based topic search

The algorithmically derived topics give a good insight into the general narratives of risks used in the documents. Nevertheless, they are hard to interpret, as early corona-related risk reports are still generic in that various risk factors are covered. Topic four, for example, provides an unspecified context with regard to the outbreak of the illness in China. Similarly, topic three covers the potential impact of the crisis in an unspecified context. In contrast, most of the business and economics related keywords appear to be covered in topic two. Moreover, the unsupervised methods are not deterministic. To ensure a robust and comparable topic identification over time, we use a dictionary-based keyword search. For this, we combine the results of the unsupervised methods with domain knowledge from economics to label the five main topics and specify defining keywords, displayed in Table A2. The most frequently mentioned bi-grams and words are considered. Using these, we can conduct dictionary-based searches in the filings. We measure the topical context of each corona-sentence by calculating the share of topic keywords relative to the word length of the sentence. This metric is later aggregated for industries and specific points in time.

Matching industry classification systems

The SEC classifies firms into industries using an amended version of the 1987 Standard Industrial Classification (SIC).⁹ The SIC is a system for classifying industries by a four-digit code. It was replaced by the North American Industry Classification System (NAICS) in 1997. The NAICS system is still in use, in its last revision from 2017, and it is being used by US government agencies. Thus, it is important to match the SIC codes used by the SEC with the NAICS system, in order to allow researcher and economists to use the CoRisk data in industry-specific applications.

To match the different classification systems, we use merging tables provided by the US Census Bureau.¹⁰ The code (*R*) and the merging tables are available on GitHub.¹¹ After having merged the four-digit SIC codes, we use the NAICS overview provided by the US Census Bureau to identify the larger two-digit sectors, on which the CoRisk data are aggregated.¹² During merging, some industries are assigned to several categories. In order to drop duplicates, we only keep the most frequent section per SIC code. While it might be possible that some information is lost during the merging process from the 1987 four-digits SIC system to 2017 two-digit NAICS sectors, we assume that less categories, which are compatible with the system used in current US statistics, provide a better overview of relevant processes than an aggregation, which is too fine-grained and which consists of low sample sizes within each category.

Methodology of the CoRisk-Index

Filing of 10-K reports to SEC

Companies with more than 10 million USD in assets or a class of equity securities that is held by more than 2,000 owners must file annual 10-K reports to the SEC, regardless of whether the securities are publicly or privately traded¹³. All 10-K reports are made publicly available by the SEC. In particular, but not exclusively, in the risk section of the report, the company lays anything that could go wrong, likely external effects, possible future failures to meet obligations, and other risks disclosed to adequately warn investors and potential investors. Companies are required to use "plain English" in describing these risk factors, avoiding overly technical jargon that would be difficult for a layperson to follow.

Comparing 10-K and 10-Q reports

In addition to annual 10-K reports, the SEC requires companies to publicly disclose their actions in a set of reporting formats (10-X). Apart from historical reporting standards and niche reporting categories, company disclosures with the SEC can be separated in two groups: Annual 10-K and quarterly 10-Q reporting. One of the four quarterly 10-Q reports is subsumed by the annual 10-K report.

We limit our analysis to the information contained in 10-K reporting alone for the following reason. On March 4, the SEC has explicitly advised public companies to assess what the coronavirus means for their future operations and financial results and to make appropriate disclosures to their shareholders and other members of the investment community. Furthermore, the SEC encouraged companies to delay SEC filings if necessary to develop the information required to make accurate and complete disclosures of the impact of the coronavirus on its operations and financial conditions. Specifically, the SEC issued an order stating that public companies that are unable, because of the coronavirus, to meet filing deadlines for SEC reports due to be filed March 1 to April 30, 2020, will have 45 additional days to file these reports so long as, among other things, they file reports on Form 8-K describing the reasons why the report may not be filed on a timely basis¹⁴. This announcement has caused an interference with the normal reporting procedures in at least two ways. Companies delayed their quarterly reporting until the last possible date (April 30, 2020) and companies were incentivised to "talk" about Covid-19 related issues.

From our perspective, this reporting bias should manifest most strongly in 10-Q reports, as they are more susceptible to short-term changes in reporting standards. We decide to include solemnly 10-K reports in our CoRisk analysis, as they have a long-term outlook of one year and a lower susceptibility to ad-hoc changes in reporting standards compared to 10-Q reports.

⁹Information about the SIC standards used by the SEC can be found here: <http://www.secinfo.com/\protect\T1\textdollar/SEC/NAICS.asp>.

¹⁰<https://www.census.gov/eos/www/naics/concordances/concordances.html>.

¹¹<https://github.com/Braesemann/CoRisk>

¹²<https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2017>.

¹³<https://www.sec.gov/fast-answers/answers-form10khtm.html>

¹⁴<https://www.sec.gov/news/press-release/2020-53>

Elements of the CoRisk-Index

The CoRisk-Index contains two statistics, (a) the text negativity, and (b) the number of 'corona'-keywords per report. Combining both measures into one index has the advantage, that the index reflects both the sentiment with which the firms report about the pandemic and the frequency they report about it.

The measure of text negativity is calculated as the share of negative words per sentence that includes at least one 'corona'-keyword. To identify negative words in SEC reports, we rely on the methodology established by Loughran and McDonald (2011)²⁰. The researchers have derived a sentiment dictionary of words explicitly applicable to the analysis to SEC filings. In addition to counting the number of negative words per sentence, we count the total number of words per sentence to calculate the share of negative words per sentence. To derive the overall text negativity measure per industry, we then calculate the average share of negative words per 'corona'-sentence per day and industry.

The number of 'corona'-keywords per report measures how often firms mention these keywords. It is a straightforward measure that is meant to assess how relevant the pandemic is for businesses. To calculate the measure, we simply count the number of 'corona'-keywords per report. While the share of firms that report about 'corona' has, in general, increased significantly during the first months of 2020, there are still substantial differences between firms and industries.

From these two measures, we derive the CoRisk-Index as the daily geometric mean for all eight industries and one for the total of all eight industries. In order to smooth out differences in the daily reporting patterns, we calculate a 14-day moving average (aligned right) of the daily CoRisk-Index values. For the online visualisation, index values are multiplied by 100 to display whole numbers.

Additionally, we provide two other measures of industry-specific corona-related risk awareness in the report and on the online dashboard: the share of firms mentioning 'corona' at least once in their 10-K reports (the count of a simple indicator variable, taking the value one, if a report contains at least one 'corona'-keyword and zero otherwise), and the share of topic-specific keywords per industry.

Representativeness

To reflect the differences in coverage between industries and to avoid excessive extrapolation to those industries that are not well represented, we use the COMPUSTAT database (made available via Wharton Research Data Services¹⁵.) to obtain employee count data from the reporting firms. Through that database, we could obtain the employee count of 4700 of 6400 reporting firms in 2020. These 4700 firms have a total of 44 million employees in 13 sectors of the US economy. The total economy has 150 million employees in these 13 sectors.¹⁶ Thus, the reporting firms represent around one third of all employees in the US economy. However, not all sectors are well represented (see Figure A5). In eight sectors, the firms filing to the SEC, for which we could obtain data, represent at least 22 % of all US employees in that sector. Therefore, we limit the CoRisk-Index to these eight industries.

The CoRisk-Index reports corona-specific risks on a daily basis. This time series nature makes the index prone to potential biases, if the industries tend to have vastly changing reporting schedules during the calendar year. We investigate the seasonal reporting pattern in Figure A6. The figure shows the share of reports filed to the SEC per quarter in the period 2014 to 2018. A clear seasonal pattern can be identified in the first quarter of each year. However, this seasonality appears to be driven by Finance. This sector reports substantially more in the first quarter. Finance has reported relatively low corona-related risks in the first quarter, despite the high number of reports. Thus, the CoRisk-Index does not seem to be biased towards those industries that file more reports in a given quarter. The other sectors do not show substantial seasonal patterns. Thus, we conclude that the results presented in the report are not substantially biased due to seasonal reporting patterns. e report are not substantially biased due to seasonal reporting patterns.

Historical robustness

The findings provided in the report rely on the assumption that the firm reports filed to the SEC actually contain information that is reflecting actual economic circumstances, and not just artefacts of specific wording, maybe due to trends, herding or currently relevant 'hot' topics. In order to investigate this, we

¹⁵<https://wrds-web.wharton.upenn.edu/wrds/support/Data/>

¹⁶<https://www.statista.com/statistics/200143/employment-in-selected-us-industries/>.

have examined historical SEC filings data²⁰¹⁷ and calculated the share of negative words (one of the two key components of the CoRisk-Index) for these filings in two ways.

First, we calculated the share of negative words per report in all 152,694 reports that have been filed between 2000 and 2018. To compare this historical text negativity with macroeconomic data, we aggregated the per-report negativity score for all reports filed in one quarter and compared this to US quarterly unemployment rates and quarter-to-quarter changes in GDP.¹⁸ Figure A7 shows the results. During the two recessions (grey bars) that happened in the period from 2000 to 2018, the GDP (upper panel) dropped significantly and the unemployment rate increased (central panel).¹⁹ Correlated with these overall macroeconomic developments, the share of negative words in 10-K reports (lower panel) increased in these periods. In periods of economic recovery (lowering unemployment rates in 2004 to 2008 and in 2010 to 2012), the overall text negativity decreased.

This observation provides supportive evidence that textual data extracted from SEC filings contains information that is correlated with the changing real-world economic circumstances. However, the overall text negativity is a highly aggregated measure, not comparable to the fine-grained data on specific topics that covered by the text analysis of sentences mentioning specific keywords, such as Covid-19. Moreover, during the course of the past 20 years, the length of 10-K reports has increased substantial, which might influence text mining measures on the report level.

Data availability

All raw data and findings are published on an interactive online dashboard. It is being made available on <http://oxford.berlin/corisk>. The dashboard provides an interactive visualisation of the main industry-specific statistics described in the report: (a) the CoRisk-Index, (b) the share of negative words per corona sentence in the panel 'Text Sentiment', (c) the share of reporting firms per industry in the panel 'Industry View', and (d) the share of topic-specific keywords per corona-sentence in the panel 'Topic Heatmap'. Additionally, we have established an easy-to-use filter and download option. It allows researchers, journalists, and the general public to get full access to both the weekly aggregated index data and the raw data on the level of individual reports.

Code availability

In addition to the online dashboard, all code, data and a detailed documentation is available on our Github page (<http://github.com/Braesemann/CoRisk>).

¹⁷<https://sraf.nd.edu/crsp-flat-file/>

¹⁸The macroeconomic data has been collected from: <https://research.stlouisfed.org/econ/mccracken/fred-databases/>.

¹⁹Recession periods from <https://www.nber.org/cycles.html>.

References

1. Fernandes, N. Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy. SSRN Scholarly Paper ID 3557504, Social Science Research Network, Rochester, NY (2020). DOI: [10.2139/ssrn.3557504](https://doi.org/10.2139/ssrn.3557504).
2. Zhang, D., Hu, M. & Ji, Q. Financial markets under the global pandemic of COVID-19. *Finance Res. Lett.* 101528, DOI: [10.1016/j.frl.2020.101528](https://doi.org/10.1016/j.frl.2020.101528) (2020).
3. Ozili, P. K. & Arun, T. Spillover of COVID-19: Impact on the Global Economy. SSRN Scholarly Paper ID 3562570, Social Science Research Network, Rochester, NY (2020). DOI: [10.2139/ssrn.3562570](https://doi.org/10.2139/ssrn.3562570).
4. Werner, E., DeBonis, M., Kane, P. & Wagner, J. House leaders look to expedite \$2.2 trillion relief package but face possibility that one GOP lawmaker may delay passage - The Washington Post. *Wash. Post* (2020).
5. Zacharakis, Z. Wirtschaftsmaßnahmen: Was das alles kostet. *Die Zeit* (2020).
6. del Rio-Chanona, R. M., Mealy, P., Pichler, A., Lafond, F. & Farmer, D. Supply and demand shocks in the COVID-19 pandemic: An industry and occupation perspective. *arXiv preprint arXiv:2004.06759* (2020).
7. Ludvigson, S. C., Ma, S. & Ng, S. Covid19 and the Macroeconomic Effects of Costly Disasters. NBER Working Paper, National Bureau of Economic Research (2020).
8. Lazer, D. et al. Computational Social Science. *Science* **323**, 721–723, DOI: [10.1126/science.1167742](https://doi.org/10.1126/science.1167742) (2009). Publisher: American Association for the Advancement of Science Section: Perspective.
9. Ramelli, S. & Wagner, A. F. Feverish Stock Price Reactions to COVID-19. SSRN Scholarly Paper ID 3550274, Social Science Research Network, Rochester, NY (2020). DOI: [10.2139/ssrn.3550274](https://doi.org/10.2139/ssrn.3550274).
10. Keogh-Brown, M. R., Smith, R. D., Edmunds, J. W. & Beutels, P. The macroeconomic impact of pandemic influenza: estimates from models of the United Kingdom, France, Belgium and The Netherlands. *The Eur. J. Heal. Econ.* **11**, 543–554, DOI: [10.1007/s10198-009-0210-1](https://doi.org/10.1007/s10198-009-0210-1) (2010).
11. Buetre, B., Kim, Y., Tran, Q. T. & Gunasekera, D. Avian Influenza: Potential Economic Impact of a Pandemic on Australia. *Aust. Commod. Forecast. Issues* **13**, 351 (2006). Publisher: Australian Bureau of Agricultural and Resource Economics.
12. Baker, S. R. et al. The Unprecedented Stock Market Impact of COVID-19. Working Paper 26945, National Bureau of Economic Research (2020). DOI: [10.3386/w26945](https://doi.org/10.3386/w26945). Series: Working Paper Series.
13. Kinne, J., Krüger, M., Lenz, D., Licht, G. & Winker, P. Corona-Pandemie betrifft Unternehmen unterschiedlich. Working Paper, ZEW – Leibniz-Zentrum für europäische Wirtschaftsforschung Mannheim GmbH (2020).
14. Goodell, J. W. & Huynh, T. L. D. Did Congress trade ahead? Considering the reaction of US industries to COVID-19. *Finance Res. Lett.* 101578, DOI: [10.1016/j.frl.2020.101578](https://doi.org/10.1016/j.frl.2020.101578) (2020).
15. Lewis, D., Mertens, K. & Stock, J. H. U.S. Economic Activity During the Early Weeks of the SARS-Cov-2 Outbreak. NBER Working Paper 26954, National Bureau of Economic Research (2020). DOI: [10.3386/w26954](https://doi.org/10.3386/w26954). Series: Working Paper Series.
16. Hassan, T. A., Hollander, S., van Lent, L. & Tahoun, A. Firm-level Exposure to Epidemic Diseases: Covid-19, SARS, and H1N1. NBER Working Paper, National Bureau of Economic Research (2020).
17. Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S. & Smith, N. A. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies -NAACL '09*, 272, DOI: [10.3115/1620754.1620794](https://doi.org/10.3115/1620754.1620794) (Association for Computational Linguistics, Boulder, Colorado, 2009).
18. Richman, L. D. et al. SEC adopts rules to modernize and simplify disclosure. *J. Invest. Compliance* **20**, 1–8, DOI: [10.1108/JOIC-04-2019-0022](https://doi.org/10.1108/JOIC-04-2019-0022) (2019). Publisher: Emerald Publishing Limited.
19. Chouliaras, A. The Pessimism Factor: SEC EDGAR Form 10-K Textual Analysis and Stock Returns. SSRN Scholarly Paper ID 2627037, Social Science Research Network, Rochester, NY (2015). DOI: [10.2139/ssrn.2627037](https://doi.org/10.2139/ssrn.2627037).

20. Loughran, T. & McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The J. Finance **66**, 35–65 (2011). Publisher: Wiley Online Library.
21. Dyer, T., Lang, M. & Stice-Lawrence, L. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. J. Account. Econ. **64**, 221–245, DOI: [10.1016/j.jacceco.2017.07.002](https://doi.org/10.1016/j.jacceco.2017.07.002) (2017).
22. Blei, D. M. Probabilistic topic models. Commun. ACM **55**, 77–84 (2012). Publisher: ACM New York, NY, USA.

Additional information

Topic number	Top 10 words
Topic 1:	impact, extent, including, outbreak, uncertain, future, highly, results, developments, depend
Topic 2:	operations, including, health, outbreak, business, supply, products, economic, public, result
Topic 3:	outbreak, spread, countries, impact, including, china, business, potential, economic, government
Topic 4:	china, outbreak, novel, covid, adversely, wuhan, strain, business, december, recent

Table 1. Results of an LDA topic modelling with four topics. The algorithmically derived topics give a good insight into the general narratives of risks used in the documents.

Topic	Keywords
Production	business operation, business disruption, product, work stoppage, labor disruption, labor, work, manufacturing operation, labor shortage, employee productivity, product development, business activity
Supply	manufacturing facility, manufacture facility, contract manufacturer, service provider, logistic provider, supply disruption, party manufacturer, supply disruption, facility, supply, transportation delay, delivery delay, supplier, business partner, supply chain, material shortage
Travel	air travel, travel, travel restriction, airline industry, travel disruption
Demand	store closure, distribution channel, market condition, consumer spend, market acceptance, consumer confidence, consumer demand, consume, store, customer, store traffic
Finance	operating result, cash flow, stock price, estate value, credit availability, performance problem
Aid	aid, coronavirus aid, care act, cares act, paycheck protection
Health	health measure, distancing, stay-at-home, governmental measure

Table 2. Keywords for five topical domains. We combine the results of the unsupervised methods with domain knowledge from economics to label the five main topics and specify defining keywords.

CBRE

Adverse economic conditions or political or regulatory uncertainty or significant public health events, such as pandemics, could also lead to a decline in leasing volume, property sales prices, funds invested in commercial real estate assets or planned development activity, which in turn could reduce the commissions and fees we earn. During 2019, our Asia Pacific business experienced declines in leasing activity amid rising geopolitical and trade uncertainty and slowing regional economies. Furthermore, in December 2019, a strain of coronavirus was reported to have surfaced in Wuhan, China, resulting in decreased economic activity in China and concerns about a potential pandemic, which would adversely affect the broader global economy. At this point, the extent to which this coronavirus may impact the global economy and our results is uncertain, but pandemics or other significant public health events, or the perception that such events may occur, could have a material adverse effect on our business.

BBX Capital

Risks that public health issues, such as the recent coronavirus outbreak, and natural disasters including hurricanes, may adversely impact the Company's financial condition and operating results, including, with respect to the recent coronavirus outbreak, impacts resulting from disruptions in the Company's supply chains, potential declines in leisure travel and consumer traffic at malls and retail locations, and public health concerns related to bulk candy products;

CORONA
TRADE
SUPPLY
TRAVEL

Figure 4. Examples of paragraphs in 10-K reports that mention the term 'coronavirus' and other business specific terms.

	Section	SEC Employees (million)	US Employees (million)	Share (%)
1.	Mining	0.7	0.75	93
2.	Manufacturing	12.5	15.7	80
3.	Information	1.7	2.8	62
4.	Wholesale & Retail	10.3	19.7	52
5.	Finance	4.2	10.8	39
6.	Transportation & Utilities	3.4	9	38
7.	Prof. & Business Services	6	19.6	31
8.	Leisure & Hospitality	3.3	14.6	22
9.	Other Services	0.4	7.6	5
10.	Education & Health	1.4	35.9	3.8
11.	Construction	0.3	11.4	2.9
12.	Agriculture	0	2.4	0
13.	Public Administration	0	7.2	0

Figure 5. Number of employees working in different sectors of the US economy and share of employees working in firms that report to the SEC. In eight out of 13 sectors, the firms that file to SEC represent at least 22 % of all employees in that sector.

Figure 6. Share of reports per quarter of the eight different industries. There is some seasonality, which is mainly driven by Finance's large number of reports in the first quarter. The other industries do not show substantial variation within a calendar year.

Figure 7. Comparison of quarterly GDP change (upper panel), unemployment rates (central panel), and share of negative words in 10-K reports (lower panel). The figure exemplifies that the sentiment of 10-K reports tend to correlate with overall macroeconomic developments.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [10k10qdynamic.png](#)
- [10kvs10q.png](#)
- [RobustnessCheckMacroNegativity.pdf](#)
- [StocksperIndustry.pdf](#)
- [approach3.png](#)
- [calendarrepresentativeness.pdf](#)
- [representativeindustries.png](#)