// NO.21-062 | 08/2021

DISCUSSION PAPER

// JULIAN OLIVER DÖRR, JAN KINNE, DAVID LENZ,
GEORG LICHT, AND PETER WINKER

An Integrated Data Framework for Policy Guidance in Times of Dynamic Economic Shocks

Leibniz Association

ZEW

An Integrated Data Framework for Policy Guidance in Times of Dynamic Economic Shocks

Julian Oliver Dörr*[1,2], Jan Kinne[1,3], David Lenz[2,3], Georg Licht[1], and Peter Winker[2]

[1]ZEW – Leibniz Centre for European Economic Research, Department of Economics of Innovation and Industrial Dynamics, Mannheim, Germany

[2]Justus Liebig University Giessen, Department of Econometrics and Statistics, Gießen, Germany

[3]istari.ai, Mannheim, Germany

June 17, 2021

Abstract

Usually, official and survey-based statistics guide policy makers in their choice of response instruments to economic crises. However, in an early phase, after a sudden and unforeseen shock has caused incalculable and fast-changing dynamics, data from traditional statistics are only available with non-negligible time delays. This leaves policy makers uncertain about how to most effectively manage their economic countermeasures to support businesses, especially when they need to respond quickly, as in the COVID-19 pandemic. Given this information deficit, we propose a framework that guides policy makers throughout all stages of an unforeseen economic shock by providing timely and reliable data as a basis to make informed decisions. We do so by combining early stage 'ad hoc' web analyses, 'follow-up' business surveys, and 'retrospective' analyses of firm outcomes. A particular focus of our framework is on assessing the early effects of the pandemic, using highly dynamic and large-scale data from corporate websites. Most notably, we show that textual references to the coronavirus pandemic published on a large sample of company websites and state-of-the-art text analysis methods allow to capture the heterogeneity of the crisis' effects at a very early stage and entail a leading indication on later movements in firm credit ratings.

Keywords: COVID-19; impact assessment; corporate sector; corporate websites; web mining; NLP

JEL: C38; C45; C55; C80; H12

*To whom correspondence should be addressed. E-mail: julian.doerr@zew.de

We are grateful for valuable comments from the participants of the Conference on New Techniques & Technologies for Statistics (NTTS2021) organized by Eurostat.

1 Introduction

COVID-19 and its economic consequences have placed numerous firms under severe distress. In almost all countries, stores and businesses were closed and mobility severely restricted to contain the spread of the virus. While these large-scale anti-contagion policies had provably positive effects on health outcomes (Hsiang et al. 2020), they fundamentally changed the landscape for many businesses. Due to the forced halt of many economic activities and the severe shock to global trade, many companies faced a situation of reduced business activity and declining sales figures, as well as major disturbances to their value chains and supplier networks, which had immediate consequences on the affected firms' financial positions.

The impact of COVID-19 on businesses has shown, however, a great degree of heterogeneity (Ding et al. 2021; Abay et al. 2020; Goolsbee & Syverson 2021). In some sectors firms have been barely affected by the pandemic or have even benefited from it, while in others large numbers of companies have been pushed into financial distress. Besides sector-specific differences, the economic exposure to the pandemic has also strongly varied with companies' business models. Some operations managed to adjust swiftly to the changed conditions, others had little scope to do so.

Now, after more than a year of pandemic, the winners and losers of the crisis seem rather clear. While firms with highly digitized business models such as delivery companies, e-commerce as well as online video conferencing and education platforms have thrived, companies whose business models are characterized by physical human interaction such as culture, travel, hospitality, restaurants and retail trade have greatly suffered (Abay et al. 2020). What seems clear today, has however not been obvious at an early stage of the shock, when policy makers were confronted with various forms of economic uncertainty (Baker et al. 2020; Njindan Iyke 2020) and stepped largely in the dark about the impacts of the pandemic on different businesses and different industries. Not only the dynamics of the pandemic were hard to foresee at an early stage of the crisis, but also governments' economic response measures have been unprecedented such that referencing to previous experiences has neither been useful nor possible. A dilemma for policy makers who were forced to act quickly to cushion the economic impact of their virus containment measures, which severely added to the plight of businesses.

In fact, the shutdown measures not only required companies to reorganize their operations by adjusting to the changed conditions, but also led to a fast erosion of equity positions among heavily exposed companies. This brought many firms on the brink of financial solvency (Dörr et al. 2021) and thus called for fast government assistance (Didier et al. 2021). Faced with the threat of a wave of corporate insolvencies and its immediate consequences such as mass lay-offs, while at the same time being confronted with an information deficit concerning the heterogeneity

of the shock's impacts, policy makers granted unprecedented liquidity subsidies and launched other support instruments[1] in a largely indiscriminate manner (Gourinchas et al. 2021).[2] The lack of early indicators signaling which firms and sectors were most exposed in the early stage after the economic shock (Didier et al. 2021) left policy makers uncertain about how to most effectively steer countermeasures. As a result, most of the early stimulus was awarded on a lump-sum basis, without taking into account that not all companies were equally affected by the pandemic. In Germany, the country our study focuses on, the government even launched the 'largest assistance package in the history of the Federal Republic of Germany' (Federal Ministry of Finance 2020b, p. 3) comprising net borrowing of the Federal Government of around €156bn (Federal Ministry of Finance 2020b). In this sense, the coronavirus pandemic has shown that grasping the economic effects of severe and sudden shocks on different sectors in a *timely* manner is crucial for policy makers to steer their response measures most effectively into directions where help is needed most urgently while not overburdening fiscal budget. This calls researchers and data scientists to explore new policy frameworks based on novel sources of data and analysis methods to guide political actors in times of sudden shocks. In this context, we believe that methods from the various fields of Data Science and the use of unstructured data bear great potential in shedding some light in situations when information deficits for policy makers are substantial and time to act is short. Therefore, we introduce the use of company website data to track the effects of the coronavirus pandemic at the firm level early on in the crisis and at a large scale. Moreover, we propose a framework that allows to guide policy makers in a timely and cost-effective manner by exploiting different sources of pandemic-related data that sheds light on the effects of the economic shock on corporations.

Usually, official and survey-based statistics guide policy makers in their choice of response instruments to crises such as the COVID-19 pandemic. However, in an early phase, after a sudden and unforeseen shock caused incalculable and fast-changing dynamics, economic data from traditional statistics is usually not yet available due to its rather inflexible and slow update cycle. This is especially true for information about smaller, unlisted companies (Fairlie 2020). Given this time delay of traditional data sources which usually guide political decision-makers, our proposed framework relies on textual references to the coronavirus pandemic published on corporate websites and state-of-the-art text analysis methods to assess the early impact of the COVID-19 shock on businesses. We refer to a coronavirus reference as self-reported text fragment (sentence or paragraph) that contains specific keywords associated to the pandemic

[1] See Didier et al. (2021) for a good overview of the various response measures that have been launched to support the corporate sector in various jurisdictions.

[2] In Germany, the country we focus on in our study, liquidity grants' 'application and payment process [needed] to be swift and free from red tape' according to the Ministry of Finance Federal Ministry of Finance (2020a, para. 2). Moreover, in context of public loan programs, 'the credit approval process [did] not involve additional credit risk assessment by the bank' and 'there [were] no requirements for collateral security' (Federal Ministry for Economic Affairs and Energy 2020, para. 5).

and the SARS-CoV 2 virus published by a company on its website. We apply our framework to a large sample containing all economically active firms in Germany that have a web domain. Our results show that our framework can provide an early indication concerning the heterogeneous effects of the pandemic on the corporate sector. In fact, it allowed us to track the impacts of the COVID-19 shock on the firm-level at high frequency and high granularity and most notably at a very early stage, way before alternative sources could reveal first patterns concerning the effects of the crisis. Acknowledging that traditional impact data such as surveys and corporate financial data provide deeper insights and more targeted policy guidance, our framework incorporates data from such sources as they become available. In our case these traditional data sources comprise results from a consecutive questionnaire-based business survey as well as proprietary credit rating data. In that sense, our framework focuses on bridging the information gap that arises when traditional data collection can only create insights with non-negligible time delays, especially in such highly dynamic situations as the COVID-19 pandemic of 2020.

The remainder of this paper is structured as follows. Section 2 provides an overview of the relevant literature. Section 3 introduces the different sources of firm-level data we use to capture the impacts of the COVID-19 shock on German businesses at different stages of the pandemic and at different levels of granularity with a special focus on heterogeneity at the industry level. The section also introduces the novel use of corporate website data for an early impact assessment after an economic shock. Section 4 empirically examines and discusses the value of company websites as early indicators of the impact of COVID-19 on the corporate sector. Section 5 concludes.

2 Related Literature

This study contributes to the fast growing literature on the economic effects of the COVID-19 pandemic. Naturally, financial markets deliver very early expectation-based insights to what extent an exogenous shock such as COVID-19 affects the corporate sector. Ding et al. (2021), for example, analyze the relationship between firm characteristics and financial market reactions using stock market information from January to May 2020 for a large number of internationally traded firms. They find that especially firms that were strongly exposed to international supply chains, with comparatively weak pre-crisis financial standing and with higher ownership by hedge funds underperformed in the months after the outbreak of the pandemic. Based on U.S. stock market returns, Ramelli and Wagner (2020) also analyze stock market performance in response to the COVID-19 shock but more strongly focus on the timing of the effects. Similarly, they find that especially internationally oriented firms that were heavily exposed to disruptions in global trade and strongly dependent on the Chinese market performed poorly especially at the

very beginning of the shock in January 2020. At a later stage, stock market reactions started to increasingly penalize firms with thin financial reserves, with consumer services deemed as the biggest losers.

Further studies based on business surveys find that firms' survival expectations shows great heterogeneity across industries and strongly depends on expectations concerning the duration of the shock's repercussions (Bartik et al. 2020). Based on a business survey conducted between March 28, 2020 and April 04, 2020, Bartik et al. (2020) find that estimated survival probabilities are particularly low in arts and entertainment, personal services, the restaurant industry and in tourism and lodging. Using the US Current Population Survey, Fairlie (2020) find that major industries such as construction, restaurants, hotels, transportation and other personal services experienced strong declines in the amount of active business owners in April 2020 due to the COVID-19 shock.

This paper also contributes to the question to which extent alternative data sources (here: foremost text data retrieved from corporate websites) and novel methods to turn this raw data into valuable information (here: methods from the field of Natural Language Processing (NLP)) may help policy makers to make informed and evidence-based decisions in otherwise uncertain environments. With increasing amounts of (often unstructured) data available, computational resources expanding and substantial advances in analytical techniques, this question has gained importance in recent years and certainly needs proof of concept. Athey (2017), for instance, argues that there are clear limits as to how 'big data' and supervised learning techniques are useful for policy guidance. This is because 'there are a number of gaps between making a prediction and making a [good] decision' (Athey 2017, p.483). The former is where data-driven models clearly thrive, the latter, however, is subject to more nuanced trade-offs which are often not encrypted in data but rather require human rationalization. Clearly, this is also true for the many policy decisions that needed to be made in response to the COVID-19 shock. Weighing between shutdown measures to contain the spread of the virus and the economic damage caused by these measures is clearly such a rationalization. Likewise, granting state aid in a whatever-it-takes fashion to prevent the risk of a wave of business failures, as well as possible windfall effects if aid measures go to non-viable firms or firms which would not have required state support, is another trade-off policy makers were confronted with in the early phase of the pandemic. While no data-driven model could have predicted the corporate outcomes resulting from different policy decisions, our study shows that the effective exploitation of non-traditional data sources in combination with data on general firm characteristics can be indicative in terms of uncovering the heterogeneity of the crisis impact on the corporate sector as well as in forecasting credit rating movements. In this sense, our framework may well serve as a guide for policy makers, especially in situations where they need to respond quickly without access to timely information

from traditional sources.

Moreover, in fragile situations where social stability is at stake, the pandemic demonstrated that it is paramount for policy makers to ensure accountability and maintain public trust in their decision making processes. Among policy makers this has led to an increasing demand for evidence-based decision making in the wake of the COVID-19 crisis (Weible et al. 2020). In this context, our framework serves as valuable evidence-based tool that allows to legitimize policy decisions in a situation of otherwise limited information.

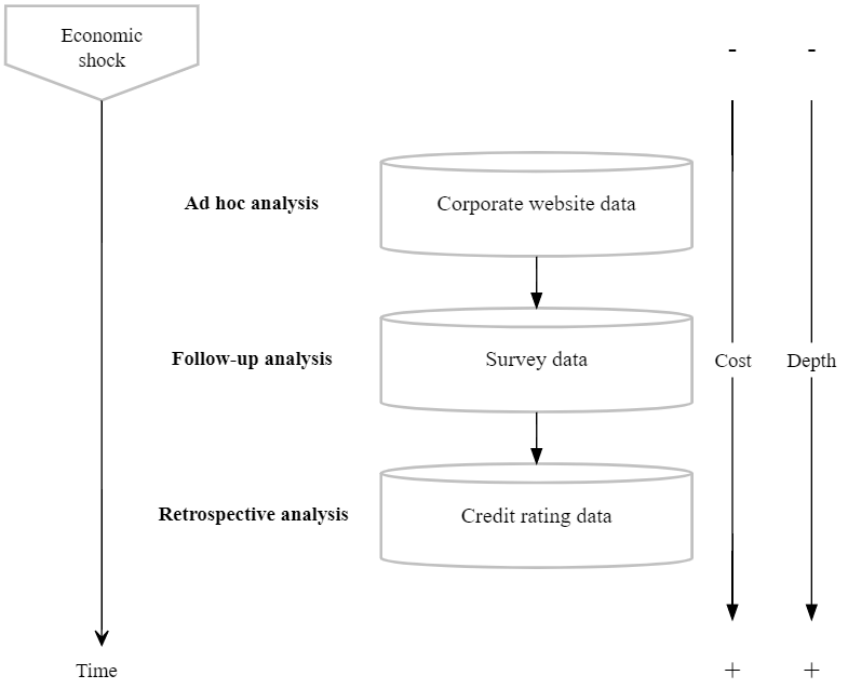
Finally, this paper contributes to the literature that exploits webdata as useful information source to tackle research and policy issues. In fact, the use of various sources of webdata to collect timely and reliable information has gained traction in recent years. For example, webdata from social media platforms is used for event detection to get an up-to-date picture of the situation regarding major social events (White & Roth 2010; Westerholt et al. 2015) or natural disasters (Resch et al. 2018; Paul S. et al. 2012). Both applications have also policy relevance in terms of public security and crisis management. In the field of economic research, data from company websites have also proven to be a valuable information resource. Companies typically use their websites to report on their products and services, to present their activities and reference customers, but also to inform their customers and partners about current events related to their business activities (Gök et al. 2015; Blazquez & Domenech 2018). Using this form of data comes, however, with a number of requirements and challenges in terms of data acquisition, data analysis and data validation. The extraction of relevant information from unstructured or semi-structured text data from corporate websites can be seen as particularly challenging here. At the same time, it promises a number of benefits, particularly in terms of granularity, timeliness, scope and cost of collection (Kinne & Axenbeck 2020). These benefits will also turn out to be key in this study. In addition to simple keyword-based approaches, e.g. to measure the diffusion of standards (Mirtsch et al. 2021), approaches with more sophisticated NLP methods in particular, have been successfully used, to generate web-based firm-level innovation indicators (Kinne & Lenz 2021), for instance. In this paper, we use corporate website data to capture and assess the dynamics of exogenous shock on the corporate sector.

In the following section, we will introduce a three-stage framework to analyze the impacts of the COVID-19 pandemic on the corporate sector in Germany. Special attention is paid to the first 'ad hoc' stage of our framework, in which we examined early phase pandemic-related dynamics on corporate websites for a large sample of German firms at 'near real time' (see also Kinne et al. (2020)).

3 Multi-stage Framework for Crisis Impact Monitoring

The framework presented in this section is based on a multi-stage process designed to provide an up-to-date and complete picture of the business situation throughout the course of an extraordinary crisis. Within each stage, different data bases at different levels of granularity are used. In a first 'ad hoc' stage, a monitoring system based on web analysis is set up in the short run, which provides reliable and up-to-date impact data at a very early phase and at 'near real time' right after an economic shock. In the second 'follow-up' stage, based on the findings of the first stage, targeted surveys are conducted using traditional methods to shed light on specific aspects of the crisis. In a final 'retrospective' stage, data collected in the aftermath of the shock or which only became available then are used to determine the impact on firm outcomes. Figure 1 gives a conceptual overview of the proposed framework and the data bases involved.

Figure 1: Framework visualization



A major advantage of this framework is that it strongly focuses on the aspect of data timeliness to ensure that policy makers are provided an empirical basis at every stage of the decision-making process. In a highly dynamic situation such as the coronavirus pandemic where policy makers are forced to react swiftly, timeliness is a key criteria. That is why alternative sources of *timely* and *reliable* data are particularly important in order to assist policy makers in designing ad hoc support measures.

In the following, we will apply this framework to analyze the impact of the COVID-19 pandemic on the corporate sector. For this purpose, we will present the data, the methods as

well as the results for all three stages of the framework.

3.1 First Stage: Ad hoc Web-based Impact Analysis

Especially in the early weeks of the pandemic after the first shutdown measures had been implemented, the impact and response of firms and in particular the heterogeneity across different economic sectors have been quite unclear until surveys and credit rating information revealed first patterns of the impact of the pandemic on corporations. We fill this information gap by making use of 'COVID-19'-related announcements found on corporate websites. For this purpose, in the first stage of our framework, we access corporate websites of about 1.18 million individual German companies from mid March 2020 to end of May 2020 twice a week and search for references related to the pandemic. Based on a labelled sample of these references, an ensemble text classifier capable of indicating in which context the company has mentioned the pandemic has been trained. This approach allows to capture first patterns regarding the effects of the economic shock on corporations and its heterogeneity across different economic sectors. In the following, we will describe how we proceeded in capturing COVID-19 references from company websites and how we turned these text fragments into a meaningful and predictive format.

In a first step, the companies' websites were queried and downloaded following a structured approach. For each corporate website address a maximum of five webpages per company (a website usually consists of several webpages) were crawled. The selection of these webpages was not conducted at random, but followed a clear heuristic: first, webpages with the shortest Uniform Resource Locator (URL) within the corporate website domain and whose content is written in German were selected (see Kinne and Axenbeck (2020) for more details on the scraping framework). The former selection criteria satisfies that those webpages with more general and up-to-date ('top-level') information were downloaded with priority making it more likely that recent Corona references are captured by the search query. The downloaded webpages were then searched for variations of the term 'COVID-19' and relevant synonyms.[3] In case of a hit, the respective HTML node was retrieved for further processing. This simple approach allowed for a first estimation of the number of companies reporting about the Corona pandemic on their websites.

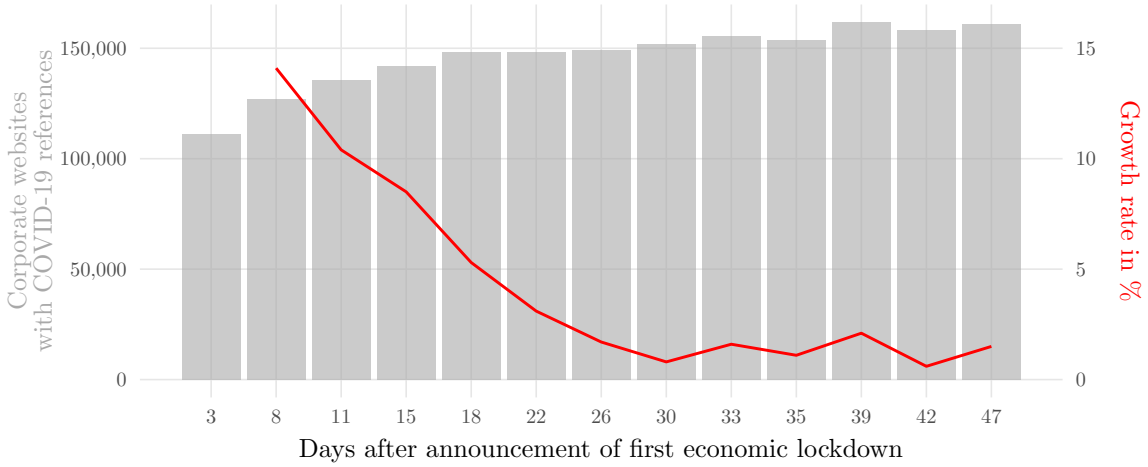
Overall, we queried the large sample of 1.18 million corporate websites 13 times over the first months of the COVID-19 crisis in Germany.[4] Figure 2 reveals that at this very early stage of the outbreak, just three days after the German Federal Government announced the first nationwide

[3] See Table 9 in the appendix for a list of these search terms.

[4] The base firm data we used stems from the Mannheim Enterprise Panel (MUP) which contains information on all economically active firms in late 2019 including the firms' web addresses (for more information on the MUP, see Bersch et al. (2014)).

economic shutdown on March 16, 2020, more than 110,000 German companies have already mentioned the term 'COVID-19' and corresponding synonyms on their websites. This comprises close to 10% of the overall corporate website addresses available to us.[5] The growth figures in Figure 2 (red line) also show that, especially at the beginning of the pandemic shortly after the first shutdown in Germany, information on company websites posed a highly dynamic source of crisis-related data, as within a few days the number of companies with Corona references grew by double digits in percentage terms. These numbers suggest that company website contents are a highly dynamic source of data in times of sudden shocks and bear great potential to learn how firms are affected by the pandemic as well as how they cope with the changed economic reality.

Figure 2: Companies with COVID-19 references on their corporate websites after announcement of first economic shutdown



Note: Figure shows the number of firms which reported about COVID-19 on its corporate website over time, shortly after the announcement of the first nationwide shutdown at March 16, 2020 (left vertical axis). The repeated design of the web queries allowed to monitor the 'near real time' impact of the pandemic on the corporate sector. Red line (right vertical axis) depicts the growth rate of companies reporting about COVID-19 on their websites. Growth rate is calculated on a rolling basis with window size 3. Fluctuations towards the last few web queries both reflect an improved scraping process that was implemented in early May 2020 and reflect companies that have removed COVID-19 references from their websites.

After the relevant text passages on company websites were identified, we proceeded, in a second step, with the classification of the context of the Corona references. For this purpose, we have made use of a pretraining multilingual language model from the Transformers family (Vaswani et al. 2017). One advantage of the Transformer model class is that relatively little training data is needed to achieve very good classification results compared to text classification models that are trained from scratch. This is because these models are based on the concept of transfer learning, which means that the models are trained on large amounts of data in a complex and computationally intensive pre-training to develop a basic understanding of language (Malte & Ratadiya 2019). These pre-trained models can then be fine-tuned on specific problems

[5]See Table 8 in the appendix for a decomposition of website addresses and detected Corona references across sectors.

and deliver very good results with comparatively few training examples. In our specific case, fine-tuning the language models involved recognizing the context of the retrieved COVID-19 references. For this purpose, we manually classified a random sample of 4,347 text passages with identified keywords into five distinct classes: (1) **Problem:** The company reports on problems related to the Corona pandemic. This includes but is not exclusive to closures of stores, cancellations and postponements of events, reports of delivery bottlenecks and short-time work (2) **No problem:** The company reports that it is not affected by the Corona pandemic or that it has no impact on its business. (3) **Adaption:** The company reports that it is adapting to the new circumstances. This includes measures such as new hygiene regulations, changed opening hours, home office and the like. (4) **Information:** The company reports generally, not necessarily in a business-context, about the Corona pandemic. This comprises general information about the spread of the virus, symptoms of the disease, news about Corona or the announcement of official regulations. (5) **Unclear:** This group includes texts that cannot be clearly assigned because they are either artefacts or the reference does not come with further clearly distinguishable content or because it is not clear for other reasons what the context of the 'Corona' designation is.[6]

Given the training data of more than 4,000 manually labelled Corona references, we then re-trained our context classification model which is based on the XLM-RoBERTa architecture (Ruder et al. 2019), a multilingual transformer model (Vaswani et al. 2017) pre-trained on over 100 languages. XLM-RoBERTa is an evolution of BERT (Devlin et al. 2018) with an improved, robust pre-training. For the final context classification, we used an ensemble method (Brown 2017), i.e. we trained multiple models with automatic hyperparameter tuning on different variants of the training data. This serves to make the overall classification more robust. The predictions of the individual models are then aggregated for the final decision and a majority decision is made.

The prediction performance of the trained model has been validated via a manual control procedure in which annotators compared the model input (text section with 'COVID-19' signal word) and model output (predicted context class) and marked them as either correctly or incorrectly classified. To do this, we had two people each validate about 450 randomly drawn model predictions. Of these 917 reviewed examples, just under 29.6% could not be labeled as predicted correctly or incorrectly beyond a reasonable doubt. This is mostly due to the fact that the cases are borderline or the identified text passages are too short for a doubtless evaluation. It should be noted here that during classification, our text analysis model receives additional text from the webpage in the form of randomly sampled single words, which is not available to the human reviewers. Overall, almost 9 out of 10 (89.5%) model predictions were classified as correct by

[6]See Table 10 in appendix for examples of each category.

Table 1: Descriptive statistics: Corporate website data

Context categories	Fraction	Mean	*N*
Problem	0.06	0.35	69,962
No problem	0.01	0.06	13,118
Adaption	0.11	0.63	128,140
Information	0.05	0.31	62,174
Unclear	0.08	0.49	98,156
Overall	0.17		202,076

Note: If a firm has reported at least one COVID-19 reference in any of the query waves that has been classified in the respective category, the firm gets assigned a 1. Else the firm gets assigned a 0 for the respective category (binarized version of the web indicators). The column 'Fraction' indicates the fraction of firms from the overall sample of 1.18 million queried websites that reported about the pandemic in the given context. Based on those firms with at least one COVID-19 reference, column 'Mean' reflects the share of firms with references in the respective context category. *N* refers to the absolute number of firms with references in the respective context category. The 'Overall' row shows the overall number of firms with at least one COVID-19 reference both in relative terms (Fraction) and absolute terms (*N*).

the human reviewers.

Table 1 provides descriptive statistics for the five COVID-19 web classes. In the table we present the classes in a binarized version where the context class for firm i equals 1 if the firm has reported on its website about COVID-19 in the respective context in any of our web queries. Otherwise the respective context class for firm i equals 0. It is worth noting that a firm can report about the coronavirus at several passages and in different contexts on its website. So the class assignments are non-exclusive. Descriptive statistics show that overall 17% of all German companies with a corporate website reported about the pandemic in some context. Moreover, 63% of the firms which reported about the coronavirus on their website, did so by mentioning adaption to the new economic circumstances. More than 1/3 (N = 69,962) of the companies with COVID-19 references signaled problems related to the pandemic and only a comparatively small number of 13,118 companies signaled the contrary of no problems. It is also noteworthy, that for almost half of the firms with a COVID-19 web reference at least one of their references could not be assigned to a narrower context group (instead, they fell into the category 'unclear').

Figure 3 provides an overview how communication about the pandemic differs across industry sectors. Most remarkably, the analysis clearly reveals disproportionately strong reporting of problems among firms in the accommodation & catering and the creative industry & entertainment sector. This clearly gives an early indication that heterogeneity of the adverse crisis impacts is substantial and that policy support in these sectors appeared most urgent. In other sectors, such as business-related services, insurance & banking and health & social services, firms relatively often informed about the pandemic on their websites which seems intuitive, especially in the latter case. Finally, it is interesting to see that companies from the insurance & banking sector relatively often signaled that they are not negatively impacted by the economic shock and that they are also strongly adapting to crisis. Deeper investigation of the references revealed that banks and insurance companies adapted to the crisis by streamlining and digitizing their

services while signaling that customer support and service quality remains unaffected by these initiatives and the pandemic in general.

Figure 3: COVID-19 firm communication on corporate websites at sector level

Note: Visualizations based on classified COVID-19 web references. If a firm has reported at least one COVID-19 reference that has been classified in the respective context class in any of the web queries, the firm gets assigned a 1. Else the firm gets assigned a 0 for the respective class (binarized version of the web indicators). Red lines represent sector-specific impact values. Grey shaded areas represent (un)-weighted average impact values across all sectors.

Table 2: Descriptive statistics: Survey data

Questions	Min	Q_1	Median	Mean	Q_3	Max	N
1 : Overall-negative-impact	0	0	1	0.77	1	1	1,478
2.A: Drop in demand	0	1	2	2.14	4	4	1,176
2.B: Temporary closing	0	0	0	1.19	2	4	1,278
2.C: Supply chain interruption	0	0	0	1.04	2	4	1,202
2.D: Staffing shortage	0	0	0	0.77	1	4	1,234
2.E: Logistical sales problems	0	0	0	0.89	2	4	1,230
2.F: Liquidity shortfalls	0	0	0	1.16	2	4	1,219

Note: Table shows descriptive statistics of survey questions. Values represent average values at the firm level across the three survey waves. Question 1 is based on a Yes-No basis. Questions 2.A - 2.F were asked on a 0-4 Lickert scale with 0 indicating no negative effects, 4 signaling strong negative effects. Non-responses in 2.A - 2.F lead to lower observation numbers in these questions.

3.2 Second Stage: Follow-up Survey-based Effect Differentiation

In a second stage, after firms have been exposed to the adverse economic environment for a critical period of time, we transfer our impact analysis from corporate website data to results obtained from a questionnaire-based survey which allows us to further differentiate the firm-level effects of the pandemic which may not be disclosed on corporate websites. For this purpose, a consecutive business survey comprising close to 1,500 distinct companies has been conducted mid of April, mid of June and end of September 2020.[7] Based on these surveys we analyze the different dimensions of the adverse impact of COVID-19 on businesses. While the survey data allowed us to capture the nature and extent of the negative impact of the current crisis on businesses in greater detail, preparation and implementation of the survey required time and resources that only allowed to obtain these insights with a non-negligible time delay after first policy measures had already been implemented. Furthermore, the information from stage 1, which is already available at an early stage, enables a more targeted design of the survey.

Table 2 shows how the design of the impact questions in the business survey allow for a deeper understanding of the various effects COVID-19 had on the corporate sector. In question 1, companies were asked on a Yes-No basis whether they are generally negative affected by the COVID-19 pandemic. For a more nuanced understanding of the type of impact of the shock and the containment measures, firms were asked in a second set of questions, in which respect they were impacted on specific dimensions. These dimensions comprise (A) drop in demand, (B) temporary closing, (C) supply chain interruptions, (D) staffing shortages, (E) logistical sales problems and (F) liquidity shortfalls and were asked on 0-4 Lickert scale.[8] Descriptive statistics of the survey results in Table 2 show that 77% of the surveyed companies reported to be negatively affected by the pandemic at least in one of the three survey waves and that a drop in demand was on average the most severe problem among the six dimensions.

[7]The survey is a representative random sample of German companies, drawn from the MUP and stratified by firm size and industry affiliation. The survey is the result of a joint research project between the polling agency KANTAR and ZEW funded by the German Federal Ministry for Economic Affairs and Energy (BMWi).

[8]0 indicates no negative effects, 4 signals strong negative effects.

Figure 4 provides an overview how the exposure to the six impact dimensions differ across industry sectors. Similar to the impact analysis via corporate website data, the survey reveals disproportionately strong impacts in accommodation & catering and creative industry & entertainment. In particular, these results allow a more precise differentiation of the negative effects, which tend not to be published by the companies on their websites and are consequently hard to detect with a web-based analysis. In particular, a sharp decline in demand and temporary closure of business operations which are associated with a liquidity squeeze have placed hotels, restaurants, catering services, libraries, museums, operator of sports, amusement and recreation facilities as well as independent artists under severe distress. The forced halt of their business activities clearly justified public liquidity support especially if the business models were running successfully before the outbreak of the pandemic. Sectors such as health & social services as well as manufacturing and engineering-related sectors show disproportionately strong exposure to the issue of supply chain interruptions and staffing shortages but are barely confronted with declining demand numbers and liquidity shocks. Clearly, for firms in these sectors policy support other than liquidity provision is required

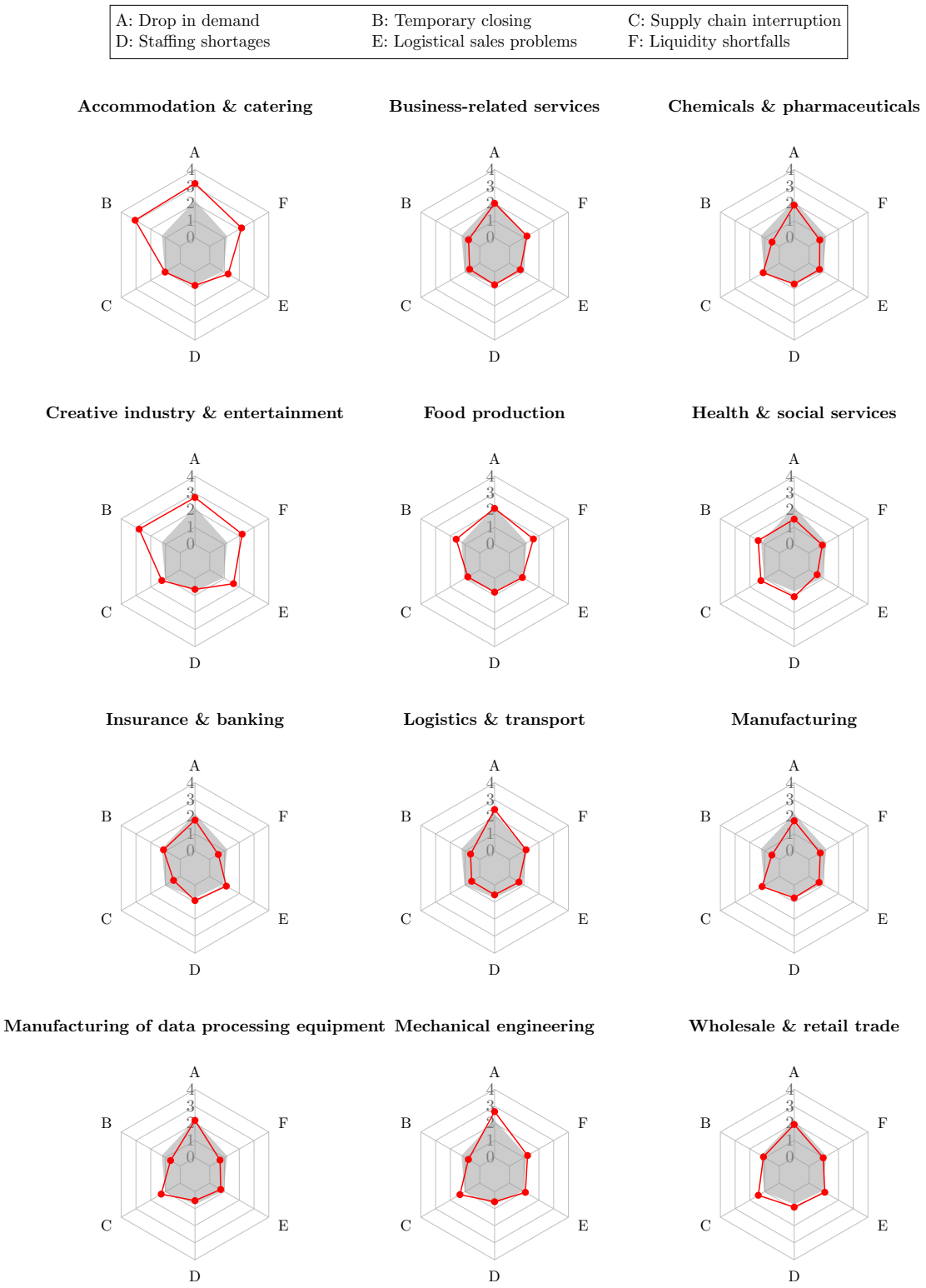
The second 'follow-up' stage of our proposed framework has clearly revealed that based on survey data especially businesses in accommodation, art and entertainment have been facing strong liquidity bottlenecks which in light of often unchanged fixed cost obligations poses a high risk of financial insolvency. In the third 'retrospective' stage of our framework, we more closely focus on this insolvency risk by analyzing the change in corporate solvency information in response to the COVID-19 crisis.

3.3 Third Stage: Retrospective Liquidation Risk Analysis

A major economic threat of COVID-19 has been and still is the risk that firms with sound business models and decent financial performance before the outbreak of the pandemic are forced into insolvency and ultimately leave the market. In a third and last stage of our framework, we focus on this liquidation risk by transferring our impact analysis from corporate website (first stage) and survey data (second stage) to firm-specific credit rating information which gives a much conciser, even though delayed[9], picture to what extent the pandemic has materialized in the firms' financial position. For this purpose, we examine credit rating updates in the crisis period for more than 870,000 German companies. While firm-specific credit rating data reflect very precise information concerning the firm's financial standing and in case of substantial credit rating downgrades signals risk of financial insolvency (Altman 1968, 2013), again the reassessment of firms' credit rating is a rather time and resource expensive process that is only

[9]We find that on average the time between two credit evaluations equals 18 months. Typically, if credit information is requested more often, a company will be re-evaluated more frequently. However, the rating capacity is largely tied to the headcount limitations of the rating agency.

Figure 4: COVID-19 firm exposure at sector level based on survey results



Note: Visualizations based on survey questions A - F which were asked on a 0-4 Lickert scale with 0 indicating no negative effects, 4 signaling strong negative effects. Red lines represent sector-specific impact values. Grey shaded areas represent (un)-weighted average impact values across all sectors. All values are averaged at the firm-level across the three survey waves.

Table 3: Descriptive statistics: Credit rating data

Variables	Min	Q_1	Median	Mean	Q_3	Max
Δr_t	-315	-3	0	3.3	0	357
Date of update	2-Jun-20	3-Sep-20	30-Oct-20	21-Oct-20	8-Dec-20	9-Apr-21

Note: Table shows descriptive statistics of the rating updates and statistics of the dates of the rating revaluations. $\Delta r_t = 0$ means that the revaluation of the company has not led to any changes in its solvency compared to the pre-crisis period. Q_1 refers to the first quartile and Q_3 to the third quartile, respectively.

available for a critical mass of companies several weeks after the shock has hit the economy.

The credit rating data that we analyze in the third stage of our framework comes from Creditreform, Germany's leading credit agency, which regularly measures and updates the creditworthiness of the near universe of active German companies. Creditreform's credit rating information is included for all firms in the Mannheim Enterprise Panel. Their corporate solvency index is based on a rich information set that closely mirrors a company's financial situation. Creditreform regularly investigates, among other things, information on the firm's payment discipline, its legal structure, credit evaluations of banks, caps in credit lines and further risk indicators based on the firm's financial accounts and incorporates this information into its rating score (Creditreform 2020b). Different weights are attached to these metrics according to their importance on determining a firm's risk of defaulting on a loan. Overall, the rating index ranges from 100 to 500 with a higher index signaling a worse financial standing.[10]

The level of the rating itself is little informative for inferring the effects of the COVID-19 crisis on the corporate sector. The change in the firms' credit rating, Δr_t, in contrast, precisely reflects to what extent a company has been down- or upgraded after the shock has hit the German economy. For that purpose, we consider all credit rating updates that have been conducted by Creditreform after June 1, 2020. We choose this date as it ensures that sufficient time has passed since the onset of the crisis to reflect COVID-related effects in the rating updates. The update in a firm's credit rating is defined as simple difference between the new rating index and the index before the update with a positive value indicating a downgrade and a negative value signaling an upgrade.[11]

$$\Delta r_t = r_t - r_{t-x}$$

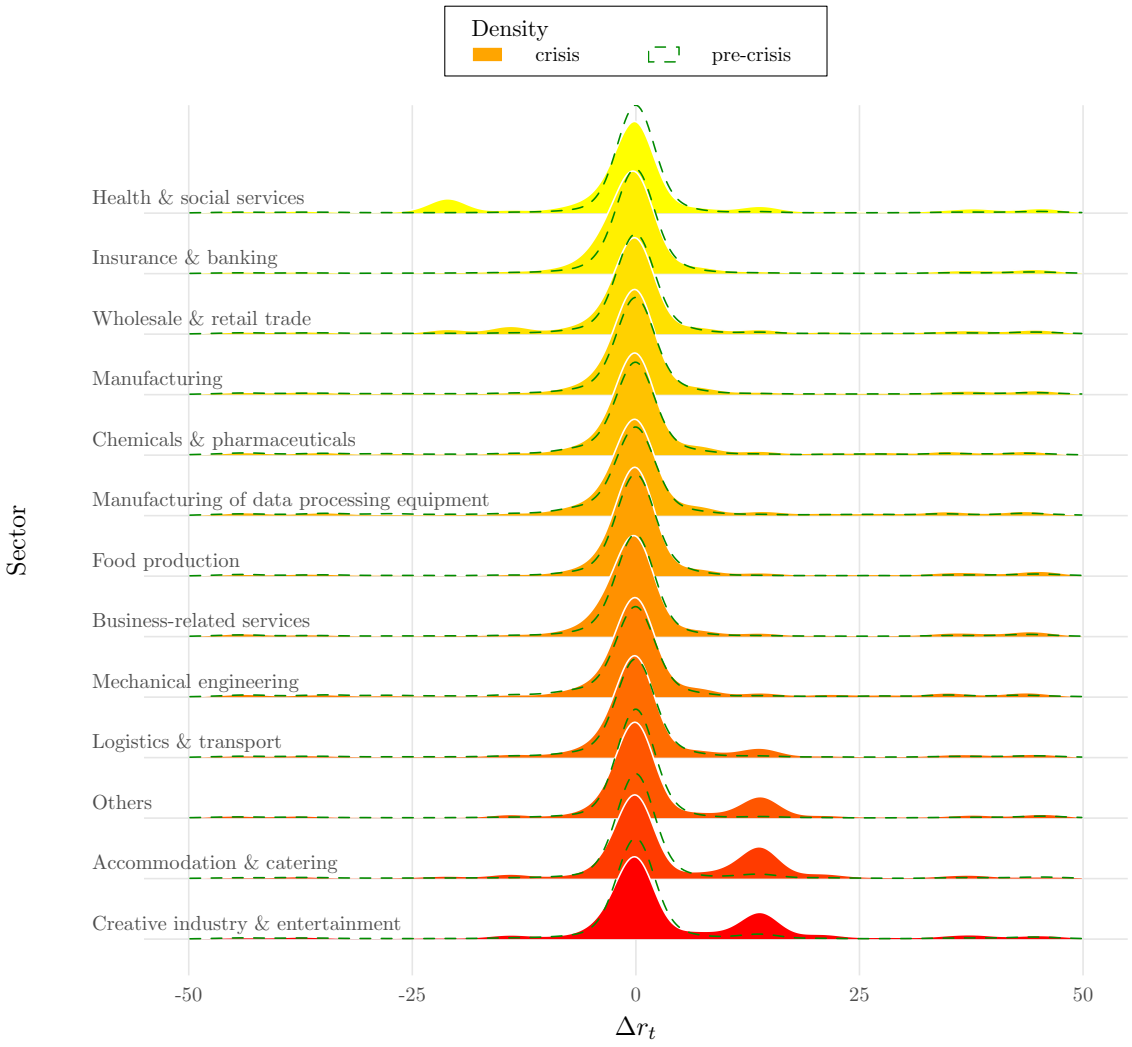
Descriptive statistics in Table 3 show that most of the distribution is centered around 0 which implies that a substantial amount of companies experienced only minor changes in their credit ratings in the COVID-19 crisis. However, taking a closer look at the distribution of the rating updates across industry sectors in Figure 5 reveals an interesting pattern: Sectors that,

[10]The credit rating index suffers a discontinuity as, in case of a 'insufficient' creditworthiness, it takes on a value of 600 (Creditreform 2020a). We truncate credit ratings of 600 to a value 500 - the worst possible rating in our analysis. We do so since our main variable of interest is the *update* in the rating index which can only be reasonably calculated if the index has continuous support.

[11]Reassessments of the rating are conducted in an irregular fashion such that the time between two updates, x, varies. On average, the time between two updates equals 18 months.

according to our first and second stage results, are severely affected such as logistics & transport, accommodation & catering and creative industry & entertainment but also supposedly winners of the crisis, most notably health & social services, follow a bimodal distribution. Comparing the crisis distribution with the pre-crisis distribution (indicated as dashed green line) suggests that this bimodality is indeed the result of the COVID-19 crisis. This means that in the crisis period larger rating downgrades and upgrades are more likely than in normal times if a sector is heavily exposed to the crisis.

Figure 5: COVID-19 effects on corporate solvency at sector level



Note: Figure shows distribution of credit rating updates both during COVID-19 (yellow to red palette) and before COVID-19 (dashed green line). Densities are based on a Gaussian smoothing kernel with a bandwidth of 2.

Ultimately, the minimum and maximum values of Δr_t show that, although rare, there are some companies that have experienced large downgrades or upgrades in their credit ratings. Table 4 shows the fraction of firms with a substantial rating downgrade of more than 50 index points within the respective sector. We see again that logistics & transport, accommodation & catering and creative industry & entertainment show a relatively high fraction of firms which experienced a substantial downgrade in their ratings compared to less affected industries as

Table 4: Distribution of extreme rating downgrades

Sector	crisis		pre-crisis	
	N	Substantial downgrades in %	N	Substantial downgrades in %
Insurance & banking	34,768	3.0	35,087	1.7
Manufacturing of data processing equipment	4,512	3.1	4,406	2.7
Chemicals & pharmaceuticals	7,204	3.1	7,000	2.4
Manufacturing	224,813	3.5	204,613	2.7
Food production	10,420	3.8	10,311	2.9
Health & social services	62,633	4.0	57,466	2.0
Mechanical engineering	12,254	4.1	12,361	2.8
Business-related services	227,957	4.3	232,576	2.4
Others	14,259	4.8	12,511	2.0
Wholesale & retail trade	173,619	4.8	169,109	2.9
Logistics & transport	39,164	5.9	37,817	3.7
Creative industry & entertainment	13,865	8.7	12,967	3.9
Accommodation & catering	44,692	9.0	36,289	4.6
Overall	870,195	4.5	832,513	2.7

Note: Table shows fraction of firms with major credit rating downgrades by industry sector in percent. Substantial downgrades are defined as credit rating downgrades of more than 50 index points ($\Delta r_t > 50$). Pre-crisis numbers refer to the year 2018.

well as compared to pre-crisis numbers. These high fractions of substantial rating downgrades reflect a relatively high insolvency risk in the respective industries. Despite the substantial policy support that these sectors received, this hints to a non-negligible number of market exits if support measures will cease before the firms have overcome the financial repercussions of the shock (Dörr et al. 2021).

4 Assessing the Predictive Quality of Early Stage Web-based Impact Indicators

The previous section has shown that all of the proposed data sources - corporate website data, survey data and credit rating data - hint to a strong degree of heterogeneity across economic sectors. While survey and credit rating data only revealed such patterns with a non-negligible time delay, COVID-19 references retrieved from company websites indicated this heterogeneity at a very early stage of the economic shock. A central question is to what extent the generated web indicators have predictive power in capturing the actual medium-term effects of the coronavirus shock. Clearly, predictive power is an important prerequisite for the web indicators to be useful for policy makers. Only if the webdata's early indication generates reliable insights, it bears the potential to help policy makers tailor their response measures and effectively channel economic assistance where it is needed most.

We assess the added value of the early web indicators by two distinct analyses: First, we compare the relationship between several firm characteristics and the negative shock exposure based on two identical regression specifications. The only difference between the two regressions

is that the we exchange the target variable which in the first regression is generated from company website information (data from the first 'ad hoc' stage) while in the second regression it stems from the business survey (data from the second 'follow-up' stage). Second, based on a sub-sample of firms for which we have both COVID-19 web references as well as credit rating updates, we analyze to what extent the classified web references serve as leading indicators for later changes in the firms' credit rating.

To examine the statistical relationships between various firm characteristics and the negative effects of the COVID-19 shock on firms, we specify a simple regression model. More precisely, we regress a binary negative impact variable on age, size and sector characteristics.

$$Probit(Y_{k,i}) = \alpha + \boldsymbol{\beta}\boldsymbol{A}_i + \boldsymbol{\gamma}\boldsymbol{S}_i + \boldsymbol{\delta}\boldsymbol{I}_i + \epsilon_i$$

with

$$Y_{k,i} = \begin{cases} \text{problem}_i, & \text{if } k = \text{Webdata} \\ \text{overall-negative-impact}_i, & \text{if } k = \text{Survey} \end{cases}$$

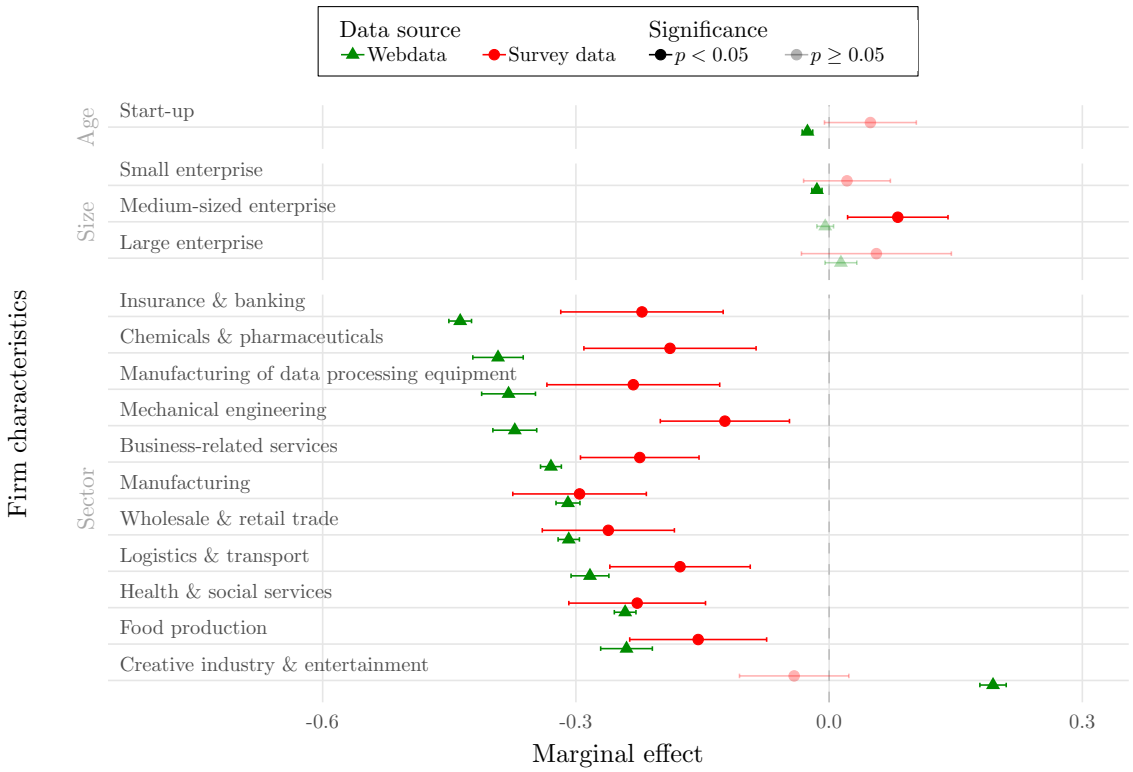
and $\boldsymbol{A}$, $\boldsymbol{S}$ and $\boldsymbol{I}$ matrices of company age, size and sector dummies, respectively.

First, we conduct the regression estimation based on the corporate website observations. The dependent negative impact variable, $Y_{Webdata,i}$ equals 1 if the firm has reported a problem on its website and 0 otherwise. Second, we estimate the same regression specification based on the survey observations where the dependent variable reflects the first question in the survey: 'Has the coronavirus pandemic had negative economic effects on your company so far?' If the firm confirmed the question, $Y_{Survey,i}$ equals 1, otherwise 0.

Figure 6 visualizes the estimation results of both Probit regressions. Effect estimates need to be interpreted relative to the reference firm which is defined as an incumbent (10 years and older), micro company (less than 10 employees) in the accommodation and catering sector. The average marginal effects thus tell by how many percentage points on average it is more likely that a firm with the respective characteristic is more likely/less likely to be affected by the pandemic. Given this interpretation of the regression results, four aspects are worth mentioning here: (i) it becomes apparent that, based on both the webdata and the survey data, age and size differences are modest at most and largely statistically insignificant in terms of their association with a negative crisis impact. However, the differences between economic sectors are substantial. Both regressions show that the probability of being negatively exposed to the shock is significantly lower in all sectors (with the exception of creative industries and entertainment) compared to the baseline sector accommodation and catering. (ii) The estimated effect directions are largely consistent between webdata and survey data and many of the estimated confidence intervals overlap. (iii) However, there are some exceptions. The most striking one being the differences

between the average marginal effect estimate for creative industry and entertainment. While the survey-based results suggest that negative effects in the creative industry and entertainment sector are statistically no more likely than in accommodation and catering, the webdata-based results hint to a significant difference between the two sectors. According to our webdata-based results, creative industry firms are more likely to be affected by the exogenous shock as indicated by an estimated gap of close to 20 percentage points. Results in Section 3.3 based on credit rating changes indeed hint to slightly more adverse impacts in the creative industry and entertainment sector relative to pre-crisis rating downgrades which suggests that the webdata effect estimate is not unreasonable. (iv) Due to the substantially higher observation number in the website-based dataset, the estimates' confidence bounds are much narrower compared to the ones of the survey estimates. The large-scale assessment that is possible with corporate website data is a clear advantage over relatively small-scale business surveys that often suffer high non-response rates. Overall, it can be said that the effects derived from webdata closely resemble the effects derived from a traditional time and resource intensive business survey.

Figure 6: Comparison webdata and survey data effect estimates



Note: Figure shows average marginal effect estimates and corresponding 95%-confidence intervals of Probit regression model where the dependent variable (negative impact) is generated from webdata (green) and survey data (red). Dependent variable from webdata reflects whether the firm has reported a 'problem' reference at its corporate website in any of the web queries. Dependent variable from survey data refers to question indicating whether the firm has suffered negative impacts due to the pandemic in any of the three survey waves. Shaded estimates signal statistically insignificant effects at the 5% level. Incumbent firms (10 years and older) serve as baseline age group, micro-enterprises (number of employees $\leq$ 10) as baseline size group, accommodation and catering serves as baseline sector among the sector dummies. Marginal effects need to be interpreted relative to the baseline group(s).

In a second analysis, we assess to which extent classified corporate website data serve as predictive indicators for later changes in a firms' credit rating. For this purpose, we regress firms' credit rating changes after June 01, 2020 on each of the five COVID-19 context classes generated in the first 'ad hoc' stage of our framework.[12]

$$\begin{aligned}\Delta r_{i,\bar{t}+z} = \alpha + \beta_1 \text{Problem}_{i,\bar{t}} + \beta_2 \text{No problem}_{i,\bar{t}} + \beta_3 \text{Adaption}_{i,\bar{t}} \\ + \beta_4 \text{Information}_{i,\bar{t}} + \beta_5 \text{Unclear}_{i,\bar{t}} + \gamma r_{i,\bar{t}-x} + \boldsymbol{\delta D}_i + \epsilon_i\end{aligned}$$

with $\boldsymbol{D}$ as matrix comprising a collection of company age, size and sector dummies.

Table 5 displays the regression estimates that result from this analysis. Regression specification (1) shows that the website categories have a significant leading indication concerning a firm's subsequent change in its credit rating. Looking at the sign estimate of the five categories, it becomes apparent that the webdata categories embody a predictive and meaningful indication concerning a firm's subsequent credit rating movement. In fact, firms which reported about problems in the context of COVID-19 on their websites suffered on average a statistically significant downgrade (positive sign) in their credit rating. Firms which have indicated that the pandemic is not causing problems on their business operations, by contrast, experienced a statistically significant upgrade on average (negative sign).[13] These results are robust when controlling for company age effects in specification (2), and additionally for size effects in specification (3). Interestingly, the statistical significance of the 'Unclear' category vanishes after controlling for company age and size which seems reasonable as the category is defined as not conveying context on the effects of the crisis. Ultimately, when controlling for sector dummies in specification (4), it turns out that even within sectors the 'problem' class has still a leading indication on credit rating downgrades as indicated by the significant positive sign estimate. The same is true for the 'information' category which still serves as significant leading indicator for later rating upgrades. For the remaining categories statistical significance vanishes when analyzing the forecasting power of the categories within sectors.

We see the results in this section as an important finding since they underpin that corporate website data serves as a leading indicator of the pandemic's financial effects on corporations. Indeed, a credit rating downgrade has typically financial consequences for a firm as it impedes

[12] $\Delta r_{i,\bar{t}+z}$ refers to the first credit rating change of firm i *after* June 01, 2020. Web categories have been extracted from corporate websites in the early crisis phase between March 2020 and May 2020, i.e. *before* June 01, 2020. We express this time period with the index $\bar{t}$. Finally, the regression incorporates the credit rating prior to the rating update which coincides with the firms' pre-crisis rating expressed via the index $\bar{t} - x$.

[13] Similarly, firms which have signaled adaption to the exogenous shock as well as such firms which only informed about COVID-19 in a broader context have also experienced upgrades on average, albeit at a lower magnitude. The same negative correlation is true if a Corona reference that could not be classified into a broader context were found on the company website.

Table 5: COVID-19 references on corporate websites as early indicators for changes in firm credit ratings

	(1) $\Delta r_{\bar{t}+z}$	(2) $\Delta r_{\bar{t}+z}$	(3) $\Delta r_{\bar{t}+z}$	(4) $\Delta r_{\bar{t}+z}$
$\text{Problem}_{\bar{t}}$	1.66***	1.68***	1.62***	0.42**
	(0.18)	(0.18)	(0.19)	(0.19)
$\text{No problem}_{\bar{t}}$	-1.70***	-1.69***	-1.73***	-0.69
	(0.42)	(0.42)	(0.43)	(0.43)
$\text{Adaption}_{\bar{t}}$	-0.46***	-0.47***	-0.33***	-0.13
	(0.08)	(0.08)	(0.08)	(0.08)
$\text{Information}_{\bar{t}}$	-0.24***	-0.24***	-0.23***	-0.17***
	(0.04)	(0.04)	(0.04)	(0.04)
$\text{Unclear}_{\bar{t}}$	-0.42***	-0.42***	-0.10	-0.08
	(0.12)	(0.12)	(0.12)	(0.12)
$r_{\bar{t}-x}$	-0.09***	-0.10***	-0.11***	-0.13***
	(< 0.01)	(< 0.01)	(< 0.01)	(< 0.01)
Age Dummies	No	Yes	Yes	Yes
Size Dummies	No	No	Yes	Yes
Sector Dummies	No	No	No	Yes
N	61,228	61,138	57,343	57,343

Note: Dependent variable, Δr_{t+1}, is the change in a firm's credit rating after June 01, 2020. Main explanatory variables of interest are the web classes generated from the website text fragments (as count variables) in the early phase of the pandemic before June 01, 2020. White robust standard errors are reported in parentheses. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

the company's ability to draw new credit lines due to its lower creditworthiness. In a phase of financial distress such as in the COVID-19 crisis, this increases the likelihood to end up in liquidity bottlenecks which may ultimately lead to financial insolvency. One problem of the sudden exogenous shock in the still ongoing COVID-19 crisis is that it has also pushed many companies with otherwise sound business models on the brink of financial solvency. From a policy perspective, this is undesirable and clearly called for quick policy support measures. In the early phase of the pandemic, the lack of information concerning the impacts on the corporate sector left policy makers little options but to grant subsidies as well as state-backed loans in a largely indiscriminate manner and at the cost of unprecedented net borrowing. Our results show that corporate website data and state-of-the art methods from the field of NLP bear the potential to cure this information deficit. With the early indication through 'ad hoc' web analyses, policy makers have a novel tool at hand that allows to detect structural distress in the economy early on. With our proposed framework, it is possible for policy makers to steer their response measures strategically to firms and sectors where help is required most urgently while not overburdening fiscal budget.

5 Conclusion

In this paper, we have presented a data-driven policy framework that not only provides policy makers with guidance for their economic support measures in times of sudden shocks, but also enables them to capture the impact of the shock on the corporate sector at an early stage. While the framework is generally applicable to assess impacts of exogenous shocks on businesses, this study focuses specifically on the repercussions of the COVID-19 pandemic on corporations. Overall, the framework consists of three stages, with each stage, according to the timeliness of the data, allows for an impact assessment at different points in the course of the crisis. These three stages, from an early stage 'ad hoc' web analysis using text fragments from company websites in the short run, to an differentiation of the various impacts via a 'follow-up' business survey in the mid-term, to 'retrospective' changes in firm's liquidity positions in the aftermath of the shock, show how information gaps that policy makers are confronted with in a highly dynamic situation such as the COVID-19 pandemic can be successfully bridged. Most notably, the early stage assessment via COVID-19 references extracted for a large sample of corporate websites is a novel and promising approach that shows how alternative sources of data and methods from the field of NLP can create insights for policy makers when traditional sources of data are only available with non-negligible time delay.

The coronavirus pandemic has shown that in situations where policy makers need to respond quickly, but information deficits make it barely possible to determine where government assistance is channeled most efficiently, public aid measures are largely granted on a lump-sum basis. In fact, in Germany, the country on which our study focuses, this information deficit has led the Ministry of Finance to choose the 'bazooka' (Financial Times 2020) instead of well-dosed and targeted liquidity injections as instrument to support companies in the crisis. Our framework is designed to help overcome information deficits which lead to otherwise undifferentiated support measures. In this context, our results show that the classification of textual COVID-19 references found on company websites allows to generate meaningful impact categories which, in turn, reveal a strong heterogeneity of the pandemic's impact at the industry level immediately after the shock and at 'near real time'. In this vein, the classified Corona references strongly resemble the exposure results which are obtained via a traditional business surveys but only several weeks after the shock has hit the economy. Moreover, we show that the classified text fragments serve as leading indicators in predicting credit rating downgrades of firms adversely affected by the economic shock. It is also noteworthy that the large-scale and 'almost real-time' evaluation via web data also enables the assessment of heterogeneity on a fine-grained regional level. While this was not the focus of this study, it nonetheless underscores the potential of the proposed framework.

There are limits to our analysis. First and foremost, not all companies have their own corporate website domain, which likely biases our web-based analysis results. Previous studies have shown that URL coverage of German companies is at 46% (Kinne & Axenbeck 2020). Especially among smaller firms the fraction without corporate URL is comparatively high. However, this does not necessarily mean that these companies do not have a corporate online presence at all. Often small and micro firms host corporate profiles on social media platforms to communicate with their stakeholders. It requires further research to detect, access and analyze these online presences to acquire an even more complete picture of corporate communication on the internet in times of economic shocks. Next, company website content is essentially self-reported information that generally bears the risk that firms communicate their current situation overly optimistic (or pessimistic). Interestingly, this study has revealed that in times of economic crises this does not seem to be necessarily the case. On the contrary, we find that close to 70,000 firms reported about problems that they are facing in the course of the pandemic. This equals 35% of all firms that published COVID-19 references on their websites and is substantial given the potential consequences of communicating 'problems' to such a broad audience.

Should machine learning-based analysis systems, such as the framework we have presented, find their way into the standard indicator toolkit of policy makers, the question of interpretable (and fair) prediction results will also arise. Complex machine learning models in particular are often deemed as difficult to understand 'black boxes' that do not allow any clear conclusions to be drawn about the factors that are ultimately decisive for predictions and forecasts. In the near future, frameworks like ours will have to integrate aspects of *explainable AI* (see for example Barredo Arrieta et al. (2020)) in order to provide decision-makers not only with reliable, but also explainable information as a basis for making informed decisions.

Despite the theoretical drawbacks of our proposed framework, we believe that it is a useful contribution to the literature on the role of data in guiding policy makers as well as a practical and ready-to-use tool. Especially in times of crisis when sudden shocks cause major disruptions, exploring alternative sources of data is crucial to be able to capture these disruptions and provide insights to decision makers in a timely manner. In this regard, we believe that webdata not only serves as a tool to capture business impacts in highly dynamic situations, but also has the potential to support policy makers across a much broader spectrum. It is left to future research to explore the value of webdata for policy at a larger scale.

References

Abay, K., Tafere, K., & Woldemichael, A. (2020). Winners and Losers from COVID-19: Global Evidence from Google Search. *World Bank Policy Research Working Paper*, *9268*.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altman, E. I. (2013). Predicting Financial Distress of Companies: Revisiting the Z-Score and ZETA® Models Background. *Handbook of Research Methods and Applications in Empirical Finance*, *1*, 428–456. https://doi.org/10.4324/9781315064277

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485. https://doi.org/10.1126/science.aal4321

Baker, S., Bloom, N., Davis, S., & Terry, S. (2020). COVID-Induced Economic Uncertainty. *NBER Working Paper*, *26983*. https://doi.org/10.3386/w26983

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., & Stanton, C. (2020). The impact of COVID-19 on small business outcomes and expectations. *Proceedings of the National Academy of Sciences*, *117*(30), 17656–17666. https://doi.org/10.1073/pnas.2006991117

Bersch, J., Gottschalk, S., Mueller, B., & Niefert, M. (2014). The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. *ZEW Discussion Paper*, *14-104*. https://doi.org/10.2139/ssrn.2548385

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*(March 2017), 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Brown, G. (2017). Ensemble Learning. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 393–402). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_252

Creditreform. (2020a). Creditreform Commercial Report International. https://www.creditreform.com/fileadmin/user_upload/CR-International/Bilder/PB_International-Commercial-Report_web.pdf. Accessed December 1, 2020.

Creditreform. (2020b). Creditreform Solvency Index. Commercial Information. https://www.creditreform.ro/fileadmin/user_upload/crefo/download_eng/commercial_information/Flyer_Solvency_Index.pdf. Accessed December 1, 2020.

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*.

Didier, T., Huneeus, F., Larrain, M., & Schmukler, S. L. (2021). Financing firms in hibernation during the COVID-19 pandemic. *Journal of Financial Stability*, *53*, 100837. https://doi.org/10.1016/j.jfs.2020.100837

Ding, W., Levine, R., Lin, C., & Xie, W. (2021). Corporate immunity to the COVID-19 pandemic. *Journal of Financial Economics*, *141*(2), 802–830. https://doi.org/10.1016/j.jfineco.2021.03.005

Dörr, J., Murmann, S., & Licht, G. (2021). Small Firms and the COVID-19 Insolvency Gap. *Small Business Economics*, *forthcoming*.

European Commission. (2003). Commission recommendation concerning the definition of micro, small and medium-sized enterprises. *Official Journal of the European Union*, *L124*, 36–41.

European Union. (2006). Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006 establishing the statistical classification of economic activities NACE Revision 2 and amending Council Regulation (EEC) No 3037/90 as well as certain EC Regula. *Official Journal of the European Union*, *L393*, 1–39.

Fairlie, R. (2020). The impact of COVID-19 on small business owners: Evidence from the first three months after widespread social-distancing restrictions. *Journal of Economics & Management Strategy*, *29*(4), 727–740. https://doi.org/10.1111/jems.12400

Federal Ministry for Economic Affairs and Energy. (2020). KfW Instant Loan for small and medium- sized enterprises to be launched tomorrow. Press Release. https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2020/20200414-kfw-instant-loan-for-small-and-medium-sized-enterprises-to-be-launched-tomorrow.html. Accessed December 29, 2020.

Federal Ministry of Finance. (2020a). Corona virus: immediate federal economic assistance now available. Press Release. https://www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2020/2020-04-01-corona-federal-economic-assistance.html. Accessed January 5, 2021.

Federal Ministry of Finance. (2020b). German Stability Programme 2020. https://www.bundesfinanzministerium.de/Content/EN/Standardartikel/Press_Room/Publications/Brochures/2020-04-17-german-stability-programme-2020.pdf?__blob=publicationFile&v=9. Accessed December 28, 2020.

Financial Times. (2020). Germany wields 'bazooka' in fight against coronavirus. https://www.ft.com/content/1b0f0324-6530-11ea-b3f3-fe4680ea68b5. Accessed June 1, 2021.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Goolsbee, A., & Syverson, C. (2021). Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020. *Journal of Public Economics*, *193*, 104311. https://doi.org/10.1016/j.jpubeco.2020.104311

Gourinchas, O. P., Kalemli-Ozcan, S., Penciakova, V., & Sander, N. (2021). COVID-19 and Small- and Medium-Sized Enterprises: A 2021 "Time Bomb"? *AEA Papers and Proceedings*, *111*, 282–286. https://doi.org/10.1257/pandp.20211109

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Huang, L. Y., Hultgren, A., Krasovich, E., Lau, P., Lee, J., Rolf, E., Tseng, J., & Wu, T. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, *584*(7820), 262–267. https://doi.org/10.1038/s41586-020-2404-8

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLoS ONE*, *16*(4 April), 1–18. https://doi.org/10.1371/journal.pone.0249071

Kinne, J., Lenz, D., Krüger, M., Licht, G., & Winker, P. (2020). Corona pandemic affects companies differently. *ZEW Short Expertise*, *20-04*. https://doi.org/10.13140/RG.2.2.11366.37441

Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *arXiv preprint arXiv:1910.07370*.

Mirtsch, M., Kinne, J., & Blind, K. (2021). Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100. https://doi.org/10.1109/TEM.2020.2977815

Njindan Iyke, B. (2020). Economic Policy Uncertainty in Times of COVID-19 Pandemic. *Asian Economics Letters*, *1*, 2–5. https://doi.org/10.46557/001c.17665

Paul S., Daniel C., & Michelle Guy. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6), 708–715. https://doi.org/10.4401/ag-5364

Ramelli, S., & Wagner, A. F. (2020). Feverish Stock Price Reactions to COVID-19. *The Review of Corporate Finance Studies*, *9*(3), 622–655. https://doi.org/10.1093/rcfs/cfaa012

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment.

Cartography and Geographic Information Science, *45*(4), 362–376. https://doi.org/10.1080/15230406.2017.1356242

Ruder, S., Søgaard, A., & Vulić, I. (2019). Unsupervised Cross-Lingual Representation Learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 31–38. https://doi.org/10.18653/v1/P19-4007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Weible, C. M., Nohrstedt, D., Cairney, P., Carter, D. P., Crow, D. A., Durnová, A. P., Heikkila, T., Ingold, K., McConnell, A., & Stone, D. (2020). COVID-19 and the policy sciences: initial reactions and perspectives. *Policy Sciences*, *53*(2), 225–241. https://doi.org/10.1007/s11077-020-09381-4

Westerholt, R., Resch, B., & Zipf, A. (2015). A local scale-sensitive indicator of spatial autocorrelation for assessing high- and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science*, *29*(5), 868–887. https://doi.org/10.1080/13658816.2014.1002499

White, J. J. D., & Roth, R. E. (2010). TwitterHitter: Geovisual Analytics for Harvesting Insight from Volunteered Geographic Information. *Proceedings of GIScience*.

Appendix

Table 6: Mapping EU NACE Revision 2 divisions to sector groups

Sectors	**Divisions**
Business-related services	58-63, 68, 69-82
Manufacturing	5-9, 12-19, 23-25, 27, 31-33, 35-39, 41-43
Wholesale & retail trade	45-47
Health & social services	86-88, 94-96
Insurance & banking	64-66
Accommodation & catering	55, 56
Logistics & transport	49-53
Creative industry & entertainment	90-93
Mechanical engineering	28-30
Food production	10, 11
Chemicals & pharmaceuticals	20-22
Manufacturing of data processing equipment	26
Others	any division not listed above

Note: Table shows the translation of EU's NACE Revision 2 divisions (European Union 2006) into the sector groupings used in this study.

Table 7: Mapping firm characteristics to size group

	Size of company			
	Micro	Small	Medium	Large
Number of employees	≤ 10	$11-49$	$50-249$	≥ 250
Annual turnover (in M €)	≤ 2	$2-10$	$10-50$	>50
Annual balance sheet total (in M €)	≤ 2	$2-10$	$10-43$	>43

Note: Table shows translation of firm characteristics into company size classes as defined by European Commission 2003 and also used in this study.

Table 8: Fraction of firms with COVID-19 references on corporate websites

Sector	**Size of company**					N
	Large	Medium	Small	Micro	Unknown	
Business-related services	53.1	38.2	25.0	12.0	8.8	356,258
Wholesale & retail trade	38.4	30.3	24.1	14.4	12.8	226,711
Manufacturing	43.1	24.9	9.7	5.0	5.3	200,021
Health & social services	82.5	59.4	42.1	25.2	24.1	166,587
Accommodation & catering	62.1	40.1	29.1	18.8	15.0	63,680
Others	89.1	75.5	62.2	32.0	27.1	40,027
Creative industry & entertainment	66.7	66.1	56.3	37.6	30.7	36,546
Insurance & banking	75.6	60.0	38.4	33.4	38.5	35,688
Logistics & transport	60.9	32.0	13.3	8.0	7.4	22,519
Mechanical engineering	45.9	24.8	10.1	5.3	5.5	11,796
Food production	20.7	18.6	13.6	10.9	9.8	11,377
Chemicals & pharmaceuticals	40.1	23.7	10.7	6.6	6.2	7,146
Manufacturing of data processing equipment	53.2	35.6	17.5	7.4	5.9	5,564
Total	59.5	38.8	23.8	14.3	15.2	1,183,920

Note: Table shows the fraction (in %) of companies within the presented sector-size strata where we could find COVID-19 references on the corporate website. Fractions reveal that larger firms are more likely to report about the virus on their websites. The numbers also show great heterogeneity across sectors. The last column presents the sample size of corporate website addresses across sectors.

Table 9: Search terms for querying COVID-19 references on corporate websites

Search terms (translated)	corona, corona virus, corona pandemic, corona crisis, covid 19, sars cov 2, wuhan virus, pandemic, 2019 ncov

Note: Searches were conducted case insensitive. Spaces in the search terms were treated as wildcards where any two characters instead of the space also led to a match. In this way, we allowed a greater degree of variation in the search for Corona references.

Table 10: Examples of COVID-19 references found on corporate websites

Categories	**Description**	**Examples** (translated)	**Class probability**
Problem	Firm reports about adverse impacts of the pandemic on its business operations.	Due to the Corona pandemic, [redacted] & [redacted] are closed.	0.98
		[redacted] has been cancelled due to the increasing concerns and escalated circumstances surrounding the recent coronavirus (COVID-19) outbreak.	0.98
		The Corona pandemic is not only affecting ongoing [redacted] projects, but also the current selection rounds of the 13th and 14th funding seasons.	0.30
No problem	Firm indicates that the pandemic has no negative impacts on its business operations.	We are there for you 24/7 as usual despite Corona!	0.86
		Your [redacted] advisor stands by your side - also in times of COVID-19.	0.75
		Corona - we are your stable partner, even in difficult times.	0.57
Adaption	Firm reports that it is adapting to the new economic circumstances.	We have also upgraded our IT and telecommunications system. Our employees are now also able to ensure that you are looked after from home, should this be necessary. Since we receive new information on the development of the coronavirus, the measures and the safety precautions every day, we will continue to monitor the development and react to it.	0.98
		Within our emergency opening times, we particularly take care of those who are currently performing at their best for our society in view of the coronavirus crisis and who depend on their glasses for their work.	0.98
		We have therefore decided to adapt our services to the current situation and to limit them until further notice. Although we want to continue to provide you with all indispensable services, we also want to meet the recommendations of the federal government on how to deal with the corona virus.	0.98
Information	Firm reports generally, not necessarily in a business-context, about the pandemic.	The corona pandemic affects each of us now and in the near future. There are many uncertainties and resulting (insurance) issues. What about entitlement to holiday cancellations, health protection abroad and coverage in the event of business interruption are just a few of the questions.	0.51
		In cooperation with the software provider [redacted], the Bundesverband Pflegemanagement (Federal Association of Care Management) is launching a platform to recruit former care professionals to cope with the currently dramatic challenges facing care against the background of the Corona crisis.	0.86
		The Association of Statutory Health Insurances has signaled support for companies that are in acute liquidity difficulties due to the Corona crisis. In particular, the interest-free deferral of contributions is massively facilitated.	0.75
Unclear	COVID-19 reference does not come with further clearly distinguishable content.	Current situation COVID-19.	0.98
		COVID-19 and how it affects us.	0.75
		Together against Corona.	0.60

Note: Table shows three website text fragments for each of the five classes retrieved from distinct corporate websites. Last column expresses the classification confidence of the text classifier expressed as class probability.

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de