Daily updated website analysis on the reaction of companies to the Coronavirus pandemic in Germany

Keywords: web analysis, coronavirus, text mining, real-time indicators, natural language processing.

1. INTRODUCTION

We analyse the websites of 1.1 million German companies twice a week for references of the word "Coronavirus" and corresponding synonyms. The identified text passages are then evaluated using AI text analysis models to determine the context of the references. This procedure allows us to assess the impact of the corona pandemic on companies in Germany in a timely and comprehensive manner.

2. METHODS

2.1. Data

This study is based on 1.1 million web addresses (URLs) of companies in Germany taken from the Mannheim Company Panel (as of the end of 2019). For each of these companies, basic company characteristics such as company location, number of employees and industry sector are also known. From previous research it is known that about half of the companies in Germany have their own website [1], with larger companies in particular having good to very good coverage (from 25 employees 84%; from 250 employees 97%). The companies' websites were queried and downloaded using an approach developed by the start-up istari.ai, with a maximum of five webpages per company (a website usually consists of several sub-webpages). The selection of these sub-webpages is not random, but follows a simple heuristic: First, sub-webpages are selected that are probably written in German and also have the shortest URL. The latter leads to the fact that mostly those sub-webpages with more general and up-to-date ("top-level") information are downloaded with priority. Overall, 81% (1.11 million) of the original 1.36 million web pages were successfully accessed and downloaded. The failed attempts are due to websites that are no longer up-to-date and permanently or temporarily deactivated.

In addition to this web data, we also have access to the results of a traditional company survey in which about 1,000 companies were asked about the impact of the coronavirus pandemic on their businesses (i.e. whether their business is negatively affected). This survey was conducted in mid-April and is available to us at the company level.

2.2. Keyword search and text analysis

The downloaded web pages were then searched for variations of the term "corona" and its synonyms (e.g. "SARS-CoV2"). In case of a hit, the affected sections of the web page (HTML markups) were marked. This simple approach allows for a first estimation of the percentage of companies reporting about the corona pandemic on their website.

In addition, the identified website sections can be further subdivided by means of computer-based text analysis, respectively the context of the mention of the "corona" term can be recorded. For this purpose, a machine learning based text analysis model specially adapted by istari.ai was used. In a first, labor-intensive step, the content (e.g. "delays in production", "employee information", "plant closures", etc.) of a random selection of

previously identified "corona" web page sections was screened and manually classified. Each web page section was classified into one of the following classes: **Problem**: The company reports on problems related to the Corona pandemic. This includes in particular closures of stores, cancellations and postponements of events, reports of delivery bottlenecks, short-time work and similar. **No Problem**: The company reports that it is not affected by the Corona pandemic or that it has no impact on its business. **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. **Information**: The company reports generally about the Corona pandemic. This includes general information such as the current spread, symptoms of the disease, news about Corona or the announcement of official regulations. **Unclear**: This group includes texts that cannot be clearly assigned because they are either artefacts or misclassifications or because it is not clear for other reasons what the context of the "Corona" designation is.

These "labelled" data were then used as training data for a transfer learning based language model [2]. During this training the model "learns" to recognize differences between the texts and is then able to automatically classify each of the identified web page sections into one of the previously defined groups. This enables an evaluation of the content of each identified website section, which provides an up-to-date picture of the impact of the corona pandemic on companies in Germany.

3. RESULTS

Figure 1 shows the development of the proportion of companies reporting on the effects of the corona pandemic on their websites. It can be seen that in March - in the first weeks of the lockdown - the proportion of companies increased much faster than in later phases. Furthermore, a methodological change in the scraping procedure is also visible, which is the main reason for the increase in the number of affected companies from 14.6% on May 2, 2020 to 15.5% on May 5, 2020.

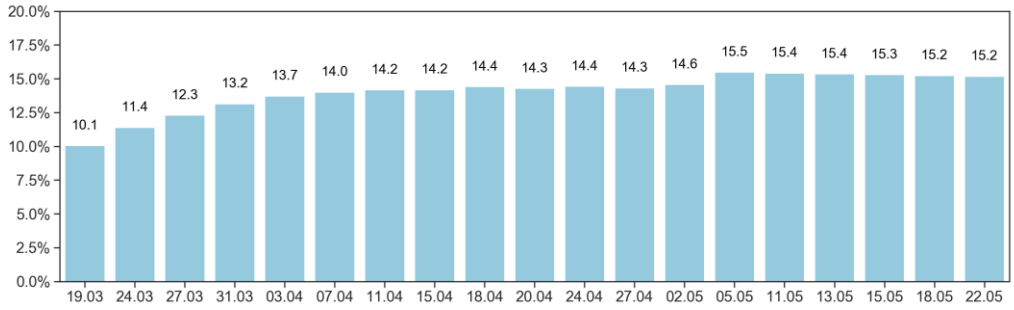

Figure 1. Share of companies (in %) mentioning “corona” on their website by date.

Figure 2 illustrates an east-west divide in terms of the proportion of companies in Germany that are regionally affected (left map). In addition, the shares of affected companies in urban districts are usually higher than in rural districts, although this difference is particularly striking in the eastern part of the Federal Republic. Regional differences can also be seen in the change over the observation period (right map). Particularly high growth rates can be observed in the western states of Baden-Württemberg, Hesse, Rhineland-Palatinate, North Rhine-Westphalia and Lower Saxony.

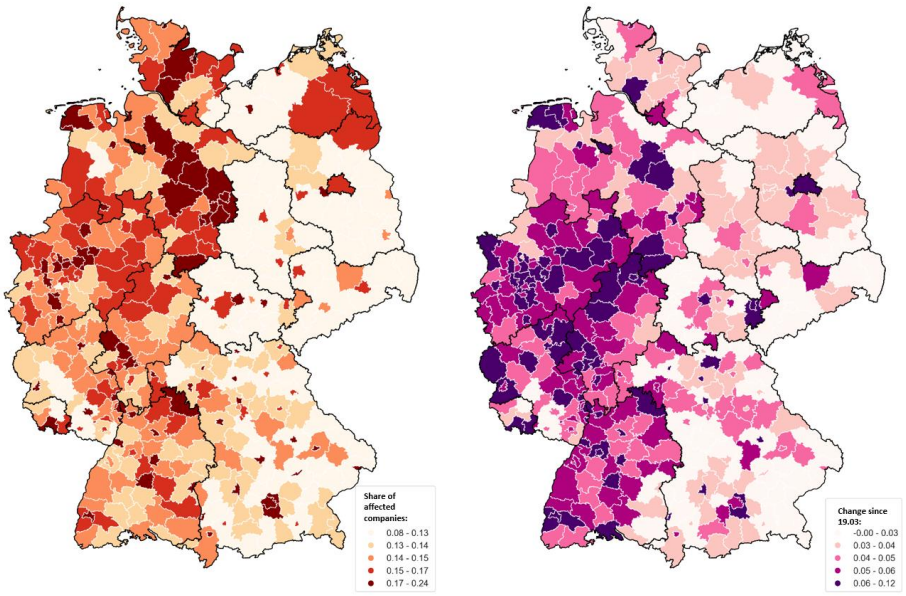

Figure 2. Share of companies (in %) mentioning "corona" on their website (left map) and change (in percentage points) over the observation period.

Figure 3 shows that companies with at least one "corona" reference are affected to varying degrees by the individual reference classes, depending on the sector. For example, public administration very often provides information on "corona"-specific issues, while the health care sector reports mainly on adjustments and the entertainment industry in particular provides information on more serious problems. The chemical and pharmaceutical, wholesale and (non-personal) services sectors appear to have no or only minor problems. It should be noted here that a company can be affected by several reference classes if several "corona" references are found on its website.

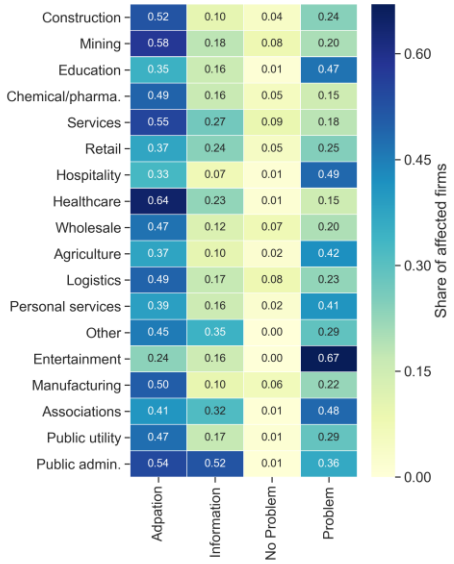

Figure 3. Type of predicted affectedness by sector, mid-April 2020.

The overlap of 109 observations between the web data we collected and the companies included in the survey is unfortunately very small. Accordingly, statistical tests in this subset usually do not provide reliable (i.e. statistically significant) conclusions. However, the survey data can be used in the context of a regression analysis to correlate the existence of a surveyed Corona-related problem of the companies with their characteristics

(company size, industry, creditworthiness, location). A repetition of this regression analysis with the significantly more extensive web data allows an estimation of whether the statistical relationships between corona exposure and company characteristics are similar in both data sets. If the results of the web-based analysis turn out to be consistent with the results of the survey, a reliable estimate of the industry-specific pandemic exposure would be possible using web analysis alone. Figure 4 shows the results of these two regression analyses as a dot-and-whisker plot in which the estimated coefficients (average marginal effects) from the logistic regressions are shown as points and the corresponding confidence intervals as lines. It becomes clear that the confidence intervals of the coefficients from the two regressions mostly overlap, i.e. can be considered consistent. In both data sets it is shown that in particular the sector affiliation is strongly related to the probability of a company having a "corona-related" problem. Interestingly, both data sets show a non-linear relationship to company size, with small companies (baseline size class up to 10 employees) being less affected than medium-sized companies, but more so than large companies with 250 or more employees.

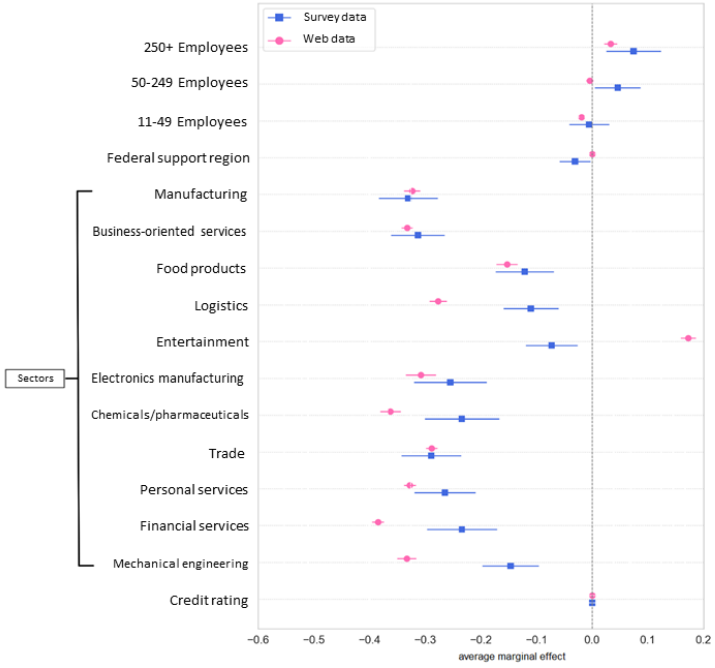

Figure 4. Regression results for survey data and web data.

4. CONCLUSIONS

The approach presented here is a first step towards a short-term monitoring of the economy and could consequently be a useful tool in steering a coordinated and evidence-based response to the current and future crisis. Since official survey data are often available only with a considerable time lag, web-based indicators can serve as an alternative and early source of information, especially in times of crisis when rapid intervention is necessary.

REFERENCES

[1] J. Kinne and J. Axenbeck, “Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany,” Mannheim, 18–033, 2018.

[2]] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018.