Investigating Businesses' Responses to COVID via Web Crawling and Text Mining

Chunming Meng
(National Institute of Economic and Social Research (NIESR), United Kingdom)
c.meng@niesr.ac.uk

John Forth
(City University of London, United Kingdom)
John.Forth@city.ac.uk

Rebecca Riley
(Kings College London, United Kingdom)
rebecca.riley@kcl.ac.uk

Paper prepared for the 37th IARIW General Conference

August 22-26, 2022

Session 7C-2, Covid-19

Time: Friday, August 26, 2022 [16:00-17:30 CEST]

Investigating Businesses' Responses to COVID via Web Crawling and Text Mining

Charlotte Meng[1], John Forth[2], Rebecca Riley[3]

09/09/2022

Abstract

Texts have attracted increasing attention over the past decade in economic analysis. To further exploit the value of texts, we collect web texts that are publicly available on a fortnightly basis for around 3,000 UK businesses operating in the UK since the outbreak of the pandemic and explore the potential to use those texts to develop timely indicators of business activities. We find that that businesses' responses are not easily identified via machine learning algorithms. Regex-based dictionary methods are more effective, but performance varies according to the action being identified. Texts from listed firms, who have the obligation to disclose business operations, have better predictive power than texts from unlisted firms. The indicators predicted by our dictionary method show that firms' actions during the pandemic are strongly correlated with listing status, revenue, employment size, and industry sector.

[1] Corresponding author. Queen Mary University of London. Email: c.meng@qmul.ac.uk
[2] City University of London. Email: John.Forth@city.ac.uk
[3] King's College London. Email: rebecca.riley@kcl.ac.uk

This research has been funded by the UKRI (ES/V004387/1) and the Office for National Statistics (ONS) as part of the research programme of the Economic Statistics Centre of Excellence (ESCoE). We are grateful for the research assistance provided by Safiya Baig, Guneek Deol, and Oona Luolakari.

1. Introduction

Data on the activities of businesses are typically collected using surveys or, in some cases, by interrogating administrative records. In the United Kingdom (UK), the Business Impact of Coronavirus Survey and the Decision Maker Panel are two examples of surveys that have provided regular information on how businesses are responding to the COVID pandemic (Gough, 2020; Mizen et al, 2022). Administrative data have also proved informative. Examples include the HR1 forms that firms must submit to the Insolvency Service when planning collective redundancies (Chiripanhura, 2021), data on firms' use of the Coronavirus Job Retention Scheme (HM Revenue and Customs, 2021) and data on the take-up of various government loan schemes (HM Treasury, 2021).

These sources are clearly of value. However, surveys are costly to administer and can suffer from low response rates, especially during crises.[4] Administrative data, for their part, are necessarily partial, can be difficult to access for research purposes and may only become available for research with a significant time lag.

In this research, we take a different approach, seeking to gather timely data on UK businesses' various responses to COVID-19 from information made publicly available on the web. Our motivations are two-fold. First, we wish to investigate the extent to which public information can replace, or complement, data obtained from traditional sources. Second, to the extent that web-based data prove to be of value, we wish to use these data to provide insights into businesses' responses to COVID through the various stages of the pandemic. Our approach is distinct from the use of web sources to generate faster indicators of economic activity because we seek to use information from the web to identify responses to the pandemic at the level of the individual firm.

To meet these objectives, we crawl the web on a fortnightly basis to collect publicly available information on around 3,500 businesses operating in the UK. Web crawling is undertaken in partnership with *glass.ai* – a UK company that has developed artificial intelligence (AI) technology to read and interpret the content from the open web, reading millions of websites and news sources. We crawl the web to search for information that describes the responses of

[4] The response rates for the first five waves of the Business Impact of Coronavirus Survey averaged 32% (Hopson, 2020). In later waves, response rates average around 25%.

our 3,500 businesses to the pandemic, across various areas of business activity, including employment, business operations and innovation.

We use text analysis to code the content of this web-crawled data and then validate the content by comparing with the results of a traditional survey of a subset of 310 firms. We find that information from these public web sources predicts firm actions to a reasonable accuracy for certain types of firms, e.g. 87% of cases for listed firms versus 53.6% for unlisted firms when considering the use of homeworking. We also identify in the data the publication bias – firms disclose actions that are likely to give them a positive image (e.g. homeworking) rather than a negative one (e.g. redundancies).

With the data and methodology described above, we further investigate firm behaviours during the pandemic. For example, using these data we find that firms that are listed and have higher revenues are more likely to adopt homeworking arrangements, make redundancies, and hire people after controlling for factors associated with prediction accuracy. Firms that are larger in size are significantly more likely to make their workers redundant.

Our research contributes to the increasing literature that uses texts as data during the pandemic (Cheema-Fox *et al.*, 2020; Hassan *et al.*, 2020; Kinne *et al.*, 2020; Dorr *et al.*, 2022; Sharma et al., 2020) by demonstrating the capability of web sources to generate firm-level data that can be used to identify and explain the behaviours of individual businesses in times of significant shocks (e.g. Brexit, COVID). Our research provides an assessment of when and when not public data might be used to substitute for or complement survey evidence and extends the use of web sources beyond conventional approaches to understanding economic activity.

The paper proceeds as follows. In Section 2, we review existing studies that have sought to investigate the impact of the pandemic on businesses, focusing in particular on prior studies that have used web crawling and textual analysis for this purpose. Section 3 then sets out the methodology of our study and Section 4 outlines the broad features of the resulting dataset. Section 5 outlines results from the textual analysis, with a focus on the relative performance of different approaches. Section 6 then reports results from the better-performing approaches, using these to investigate the types of firms that responded in specific ways to the pandemic.

2. Related literature

A growing body of literature seeks to analyse the economics effects of the COVID pandemic on businesses in the UK. Much of the existing evidence for the UK relies on surveys of businesses. The Business Impact of Coronovirus Survey (BICS) run by the Office for National Statistics (ONS) is a voluntary, fortnightly survey which investigates changes in businesses' trading status, financial performance and working arrangements during the pandemic (see Hopson, 2020; Gough, 2020). The survey indicates that around one-third of businesses ceased trading in the early months of the pandemic. Around one half experienced lower turnover than they would normally expect for the time of year. Experiences varied considerably across industry sectors, however, with businesses in the hospitality and leisure industries most likely to experience adverse effects. BICS also shows that a minority of businesses, including some wholesale/retail businesses, saw an increase in trade. Similar insights come from the Decision Maker Panel (DMP), which also highlights reductions in capital expenditure and extensive use of home-working (see Mizen et al, 2022). The DMP is a monthly survey of around 3,000 chief financial officers of small, medium and large firms in the UK, run partly by the Bank of England. The DMP differs from BICS in that many of its questions ask respondents explicitly to *estimate the impact of COVID on business activity*, relative to what would otherwise have happened. Other surveys of businesses include that by the Centre for Economic Performance (CEP) and Confederation of British Industry (CBI) on technology adoption, which indicates that around three-quarters of CBI member organisations adopted digital technologies in the first year of the pandemic (Valero et al, 2021). Similar surveys have been undertaken in other countries (e.g. Bartik et al, 2020; Garcia et al, 2020; Bellman et al, 2022)

The key focus of the current paper is whether (or when) one can use online text and methods of textual analysis to identify and explain the behaviours of individual businesses during COVID. A number of existing studies have sought to use publicly-available text to analyse the experiences of individual firms. Kinne et al (2020) analyse the websites of around 1.2m German companies twice a week from March 2020 to May 2020, searching for references to the pandemic via a set of COVID-related keywords. Relevant text passages are classified into one of five context categories (problem; no problem; adaption; information; and unclear) using a pre-trained language model (see also Dorr et al, 2022). The results indicate the broad nature of each businesses' experience with high-frequency. In keeping with survey evidence, they reveal strong heterogeneity by industry sector, with strong adverse impacts in hospitality and leisure industries. Lee (2020) also reports on efforts by the Office for National Statistics in the

UK to analyse business websites. In this case, the websites of around 0.5m businesses are analysed using natural language processing in an attempt to identify whether the business' trading had been interrupted by COVID. Again, they use COVID-related keywords to identify relevant text, but the challenges of interpreting the text content mean that few inferences are drawn from the analysis.

Hassan et al (2020) analyse data from the transcripts of earnings calls held with investors by around 12,000 firms worldwide in the first months of the pandemic. They use COVID-related keywords to identify text passages relating to the pandemic, subsequently classifying texts according to 'disease sentiment' (positive or negative tone) and 'disease risk' (indicating firm risk or uncertainty). Their data is relatively low-frequency, since most firms hold earnings calls on a six-monthly basis, but has the advantage that managers are obliged to reveal material issues to investors; the calls also provide space for questions, encouraging managers to comment on things which they might otherwise have chosen to ignore. They show that, in the early phase of the pandemic, firms commonly perceived the main risks to relate to a collapse of demand and disruption in supply chains, with other perceived risks relating to capacity reductions, site closures, and employee welfare. Sharma et al (2020) also use text announcements from firms but, in their case, rely on information distributed by NASDAQ 100 firms via Twitter between February 2020 and May 2020. The specific focus of their research is supply chain problems. From an analysis of frequently-appearing n-grams, they are able to identify a number of recurring themes, including demand surges and a lack of technological readiness. Unlike Hassan et al. (2020), however, they do not seek to investigate the firm-level correlates of these problems.

Whilst the aforementioned studies focus on firm-generated content, Cheema-Fox et al (2020) study companies' responses to COVID by studying news coverage of around 3,000 large, listed companies from around the world. They search news articles which name their sampled firms and use natural language processing to classify the subject of the news item (supply chain; human capital; or products and services) and to analyse the sentiment of each news report. They find that companies with more positive news (e.g. avoiding layoffs) have less negative stock returns, but the elasticity varies by type of company and country.

The literature on text analysis identifies a number of generic challenges when seeking to discern meaning from text (see Gentzkow et al, 2019; Grimmer and Stewart, 2013). A number

of these challenges are evident in the previous studies discussed above. For instance, individual words take on different meanings depending upon the context in which they are located, and so it is often necessary to go beyond the "bag of words" approach. The nuanced use of negation in natural language (e.g. we chose not to close our premises) may mean that it may be difficult to discern the sentiment of a sentence, and thus whether a problem has been encountered or avoided.

The existing literature also reveals a number of challenges which are particularly evident when attempting to identify businesses' responses to COVID. In any firm-generated data, there are likely to be reporting biases arising from companies' desire to avoid attracting negative publicity or, more generally, from the general public's need for information about the status of the business (Kinnie et al, 2020; Lee, 2020). There are also likely to be challenges in determining whether some action has necessarily been caused by the COVID pandemic or is simply coincident with it. Identifying relevant text is also a challenge, since the COVID context may sometimes be assumed rather than being explicitly stated (Lee, 2020).

Our study contributes to the existing literature by seeking to use publicly-available text to classify the experiences and actions of a representative sample of UK firms during the COVID pandemic. We explicitly investigate the extent to which public information can replace, or complement, data obtained from traditional sources, thus contributing to the methodological literature on text analysis of business data. Where robust insights are found to be possible, we then use the data to analyse businesses' responses to COVID through the various stages of the pandemic, thus contributing to the literature on the economic effects of the disease.

3. Methodology used for data collection

3.1 The population and sample of firms

The population for our research comprises all companies with 51 or more employees, operating in SIC (2007) Sections A-S. ONS data indicate that firms with 51 or more employees account for 50% of all employment (61% of private sector employment) and 66% of all output (70% of private sector output).

To construct a sample that is representative of this population, we drew a stratified random sample of 4,135 firms from the 32,026 in-scope firms recorded in the FAME database of company accounts as of June 2020. We first sought to sample all 296 firms that participated in

the 2014 World Management Survey; the aim was to facilitate later analysis using the data collected in the WMS. Larger firms were then over-sampled, in recognition of their disproportionate contribution to overall employment and output. Firms listed on the stock market were also over-sampled, in recognition that they have legal requirements to publish information on their activities to shareholders.

Upon drawing the sample, we extracted a number of fields from the FAME database, including company name, registered number, primary UK SIC2007 code, website, listing status, contact information (postcode, email address and phone number), and key statistics in 2018 (turnover, number of employees, fixed assets).

Glass.ai sought to validate the company website information provided by FAME, updating this information where appropriate. A number of deletions were made from the original selection of 4,135 firms as part of this process. Some 372 firms were removed due to no website being discovered. A further 101 firms were removed because the website listed in FAME was found to be inactive or because of poor evidence that this was the correct website, with no better information emerging from Glass.ai's own searches. Finally, 173 firms were removed as they were found to share a website with another firm in the sample (typically, a parent company).

The final sample consisted of 3,489 firms. Some 138 of these firms featured in the 2014 WMS. Some 830 of the 3,489 are listed firms. Table A1.1 in Appendix 1 presents the full sample distribution, whilst Tables A1.2 and A1.3 provide further information on the breakdown of this sample and of the associated population.

The use of variable probabilities of selection means that the selected sample will not be fully representative of the population. To remove this sampling bias, we constructed weights that restore the profile of the sample to that of the population across the specific dimensions used in the sampling scheme (i.e. WMS status, employment size and listed status). These weights are utilized when seeking to use the sample of firms to infer the behavior of the population that the sample is intended to represent.

3.2 Sources of public information

We collect public information for each of our sampled companies from three distinct sources, i.e. text appearing on the company website (including an investor-relations micro-site), publicly-available company reports (PDF documents published on the company website or on

an investor-relations micro-site), and news items published by online news outlets such as BBC. Section 3.4 provides further information on these sources and how data are collected from each one.

3.3 Identifying COVID actions

Our focus is actions that businesses have taken in response to the COVID pandemic. In order to identify COVID-related content in the three web sources mentioned above, we define a set of COVID-related key words (search terms) (see Table 1). We only collect public information on our sampled businesses in cases where the text packet (a web page, company report or news item) contains one or more of these COVID key words.

Table 1. Search terms for web crawl

Search terms
Exclusive terms:
Coronavirus / "corona virus"/corona-virus
Covid / covid19 / covid-19
SARS-Cov-2
Omicron*

* Added to the list on 17th December 2021

3.4 Web crawl

Glass.ai developed a web-crawling protocol that searches the three sources mentioned above (websites, news and company reports) for mentions of any of the COVID keywords listed in Table 1. The process is, however, slightly different depending on the source.

For company websites, we search the corporate website of each of our 3,489 sampled firms daily, searching for direct and indirect COVID terms. Some 255 of these 3,489 firms have separate 'investor relations' micro-sites and these are also included, giving a total of 3,744 websites. If any page on any one of these websites contains one of our COVID search terms, the content of webpage is extracted to our dataset. If the content of the webpage is altered subsequently, it may be crawled again at a later date, even if the specific portion of text containing the COVID key words remains unchanged.

In some cases, companies provide information via their website in PDF format rather than as web text. This is often the case for press releases or annual reports. Once each fortnight, the web crawler searches for PDF documents on the company website. It searches the text inside those PDFs for any mention of our COVID search terms. If a PDF document contains one of these terms, the content is downloaded in text format.

For news sources, we search 767 national and local news websites, either by crawling their website directly or by searching via Bing. We do this daily, searching for news items that mention any one of our COVID search terms and any one of our sampled firms within the same news item. If a match if found, the full text of the news item is extracted to our dataset. The web crawler searches for company names using a deterministic match to the name listed on FAME, after omitting generic suffixes (e.g. Company, Ltd, PLC). Exceptions are made for companies for which the non-suffixed name is very short (five characters or less) or equates to a dictionary word (e.g. Bottle Co.); in these cases the suffix is retained to avoid false positives. In some cases, multiple companies from our sample are mentioned in the news report; in these instances, the text packet is duplicated in our dataset so that an entry is made for each sampled firm.

Extracted text is stored in our dataset against the company reference number (CRN) for the sampled firm. Other saved fields include the URL, the search terms that have triggered the extract and the date that the text was published (where known); see Appendix 2 for a full set of data fields.

3.5 Pre-processing

The collected texts must pass some initial screening in order to be included in the data. They must be written in English and have minimum length of 50 characters. Some firm websites put a notice in the header or footer of their website pages with a sentence such as "See our measures to tackle coronavirus here." This notice may appear on every page on their websites (including those without any further details of the company's actions in relation to COVID) and, in these cases, has the potential to generate many false positives. We screen out such notices as they contain no useful information.

Some documents are very lengthy and, where this is the case, it is likely that the document covers multiple topics, some of which may not be related to COVID. To focus on the most-relevant text, we identify the location of each COVID keyword in the text, and extract 100

words before and after this COVID keyword. This 200-word text extract then forms the "text packet" for the purposes of our analysis. Under this approach, it is possible for one news report (say) to generate multiple text packets.

3.6 Time frame

Glass.ai began crawling the web on 8th July 2020 and have continued through to 15th July 2022, although the data analysed in this paper only cover the period until 2 June 2022. Necessarily, the web crawler could only cover material appearing on the web from 8th July onwards; however, any content published on the website at an earlier point in time and still present was captured at this point.

Some revisions were made to the data collection protocol in the early weeks of web-crawling, primarily to narrow the scope of the data collection to avoid obvious false positives (see Section 6 for examples). All news sources were also crawled again in early August 2020 to search for all historic news content published since January 2020. The process was completed, and the final crawling protocol was fixed, in mid-August. A dataset aligned with this protocol was was provided to ESCoE on 27th August 2020, representing the first official data packet. Data have been supplied on a fortnightly basis since that time.

One further substantive change has been made since 27th August: an expansion of the scope of the PDF document collection in the first half of September 2020. In August, the web crawler focused its search for PDF documents on webpages or micro-sites specifically dedicated to investor-relations. However, it became apparent that, occasionally, companies would publish pertinent PDF documents elsewhere on their websites: examples include press releases and COVID notices for customers. The web crawling protocol was expanded on 10th September to search across the whole of the company website for relevant PDF documents. Any documents published prior to this date and still appearing on the company website at that time were collected and appear in the dataset supplied on 10th September. PDF documents obtained from webpages or micro-sites specifically dedicated to investor-relations can be identified in the dataset via the field '*ir_filter*'.

3.7 Filters to limit advice and commentary

Firms in certain industries posts news, advice and commentary on their websites. Our keyword-searching web crawl will pick up all contents about coronavirus from those websites. However,

most contents are not about actions of this specific firm, but about other firms, government policies or the general economy. For example, some firms provide consultancy services to businesses. They may, for example, post articles about government funds to support businesses since the pandemic. Such articles are not useful for us to parse information on the firm's own actions to handle the pandemic. More importantly, they lead to false indicators in categorisation. We identify a set of industries where firms commonly use their websites to disseminate general news, advice and commentary, based on inspection of our text corpus. We then exclude firms from these industry sectors from our analysis. The industry sectors that we excluded (and their associated UK SIC(2007) codes) are as follows:

58.13 – Publishing of newspapers
60.10 – Radio broadcasting
60.20 – TV programming and broadcasting
69.10 – Legal activities
69.20 – Accounting, auditing and tax consultancy
70.2 – Management consultancy
73.20 – Market research
84.11 – General public admin (regulation)

The exclusion affected a total of 113 firms. Accordingly, after the exclusion, the sample contains a total of 3,376 firms. The distribution of the remaining firms is presented in Appendix 3.

3.8 Ethical considerations

Collecting data from company websites comes with some ethical considerations. We conduct our web crawl in a way that is lawful and which does not infringe intellectual property rights, website terms and conditions or potentially fall foul of the Computer Misuse Act 1990 or data protection legislation. In addition, we honour any requests made by website owners to refrain from reading their publicly open website and follow web etiquette (e.g. obeying the robots.txt file).

Finally, we do not re-publish any open web data that may identify individuals. Any sharing of identifiable open web data with researchers outside the project is [would be] covered by a data sharing agreement that requires those researchers to maintain the anonymity of individual persons.

3.9 IFS surveys

In order to provide alternative information on the actions of our sampled firms, we ran a short business survey in which we commissioned a research company with extensive experience of business surveys to interview businesses about their responses to the pandemic. We did not expect all firms to respond: only 311 did so. However, this provides information on the types of firms that are most likely to be missing from sources based on surveys. In cases, where the business did respond to the survey, we obtained a direct report from a senior role holder within the firm as to how it acted over the period in question. Direct comparison of the actions categorized from the text with the actions reported in the survey for the same firms therefore provides one means of validating the results of the text analysis.[5]

On the issue of non-response, we expect that certain type of firms are more likely to respond to surveys during crises, when they are dealing with tough business situations. For example, larger firms are likely to have the ability to navigate and respond while small firms may be busy "fire-fighting". To understand the selection into the survey, we regress a dummy that represents whether a firm has responded or not on a number of firm characteristics, including log turnover, whether a firm is listed or not, employment size and industry dummies. The results can be found in the Appendix 4. Listed firms, firms with higher revenues, and firms engaged in business services and public services were more likely to respond to the survey than other types of firms.

4. Dataset contents

Appendix 5 provides some descriptive statistics on the number of documents collected over the period July 2020 to June 2022. To this point, we have collected a total of 1.6m COVID-related documents, comprising 520,000 website pages, 470,000 news items and 607,000 report extracts. Among the 3,489 firms in our sample, 3,256 have had at least one COVID-related document read from the web.

It should be noted that the documents listed in Appendix 5 are counted prior to any validation on their content. Many will not contain information on the business actions that are the focus of the project; the process of identifying relevant texts within the corpus is described below.

[5] This does, however, assume that the survey responses are accurate.

5. Textual analysis

We use textual-analysis methods to classify each document in our dataset according to whether the document records some aspects of business activities in relation to the pandemic. Activities might include: closing a work site; shifting the workforce to remote working; introducing a new product or service; experiencing a drop in revenue etc.

We take two approaches to text classification. We first apply the N-gram method to obtain an overview of the contents of the texts and most-commonly-discussed firm actions. Then we use a refined dictionary method to investigate three specific actions, i.e. home-working arrangements, making redundancies, and hiring.

5.1 N-gram method

An N-gram is a portion of a text string, N words in length. A unigram is one word in length, a bigram is two words in length, and so on. In the string "Here comes the fox", there are four unigrams (single-word components) and three bigrams (two-word components).

We use Python to process each document in the dataset and identify the top 10,000 commonly-occurring unigrams and bigrams. From the resulting list, we are able to identify frequently-cited terms that appear in the documents contained within the dataset. We identify the most-frequently-mentioned terms that seem to relate to business activities, and use these to identify eight categories of business activity that are commonly mentioned:

- Business continuity
- Risk management
- Products and services
- Technology and technological innovation
- Home-working
- Health and safety
- Supply chains
- Financial performance

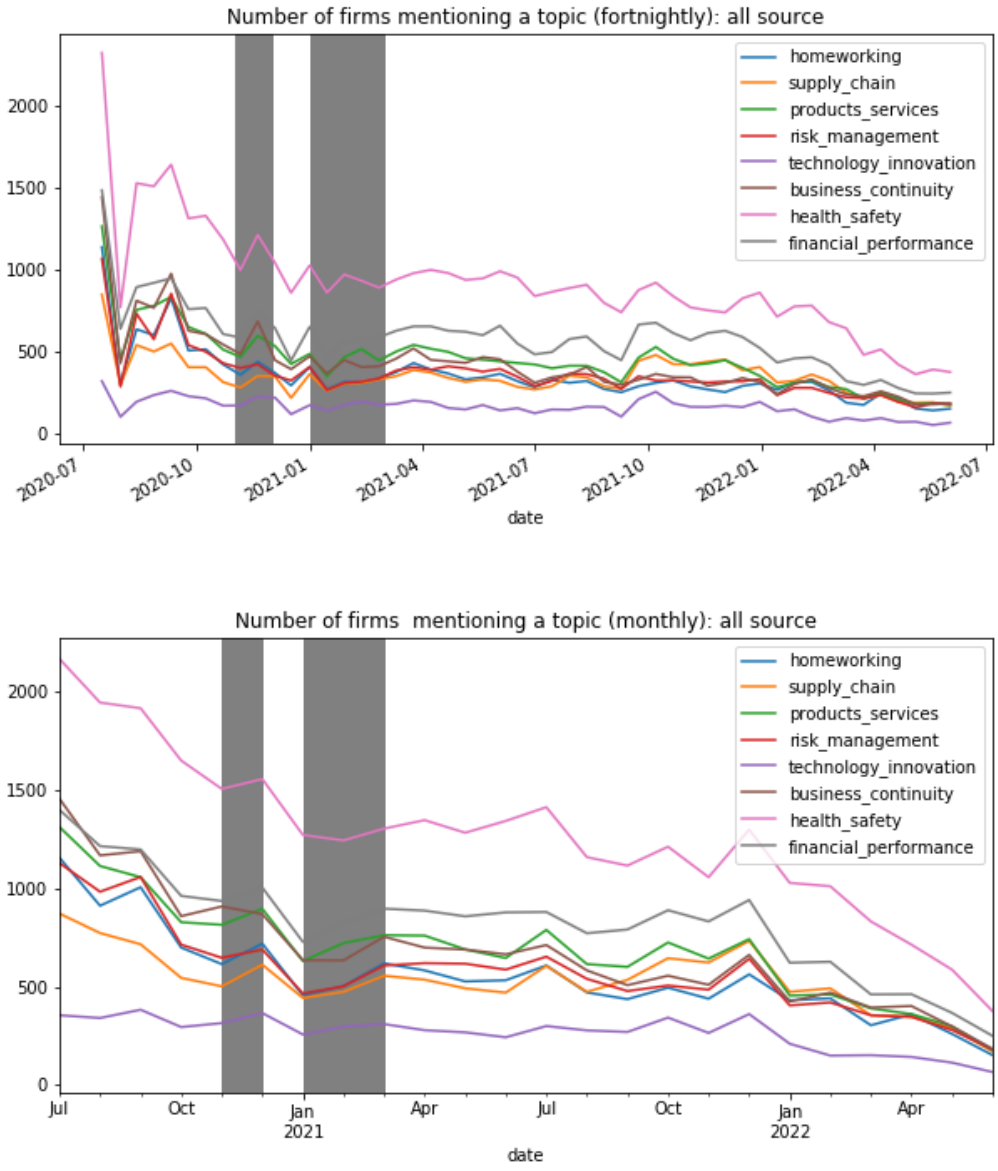
We then go through the top 1,000 bigrams to compile a list of terms that occur under the above eight categories. For example, there are seven bigrams used to describe homeworking-related

actions, i.e. *home work, remot work, return work, work home, work remot, flexibl work*, and *hybrid work*[6]. The full list can be found in the Appendix 6. We then use the list to tag the documents and determine if something has been published about the firm in relation to a certain business activity in a certain period.

Figure 1 presents the frequencies of business activity mentions from July 2020 to June 2022. The shaded areas show the two national lockdown periods in the UK, i.e. Nov 2020 – Dec 2020 and Jan 2021– March 2021. During the lockdowns, all topics had an increased coverage. "Health and safety" is the most frequently discussed topic throughout the period examined. "Supply chain" and "Product and services" experienced increased coverage from July to December 2021.

[6] We follow the common practices in natural language processing and have pre-processed the texts before applying the N-gram technique by removing stop words, removing numbers and punctuations, stemming and lemmatization. As the result, the terms are not the original words anymore.

Figure 1. Frequencies of business activity mentions over time (July 2020 – June 2022)



Notes: The shaded areas show the two national lockdown periods in the UK, i.e. Nov 2020 – Dec 2020 and Jan 2021– March 2021. Our data do not capture the first lockdown period, i.e. March 2020 – June 2020.

Two problems arise when we use the bi-grams. Firstly, we cannot determine whether the text is just commenting on an issue or reporting on the firm's activities. For instance, a financial report on the website of a restaurant chain may refer to changes in the societal extent of homeworking as an explanatory factor, rather than reporting on the extent of homeworking within the business itself. Secondly, we cannot identify from a simple bi-gram whether a firm is starting or planning to take an action, or whether it is stopping or avoiding this action.

Therefore, having identified these eight broad topics as the ones that are most frequently mentioned, we train a machine learning (ML) algorithm to automatically classify documents on the basis of whether the document mentions one or other of these areas of firm activity. To train the model, we employ a training dataset. This is a subset of documents that have been manually coded to indicate whether they contain the activity of interest. In the coding process, we specifically tag documents where the text: (a) relates to the activities of the firm in question; and (b) relates to the firm starting or stopping an activity. If we generate a well-functioning algorithm, this can then be used to classify the remaining documents in the dataset, and this classification can then be used to describe firms' behaviours through the pandemic, using all documents in the dataset.

However, we encountered with two challenges when constructing the training dataset. First, the sample is extremely imbalanced. Even for the most frequent topics, the share of all texts that mention a firm having taken an action within this domain is still very low (see the first row of Table 2). For example, the share of texts mentioning a firm having made an action relating to health and safety is 0.07 (7 per cent), meaning that for each 1000 text packets examined – which takes a coder around 16 hours to manually tag across all eight actions – only 70 cases are coded one and 930 are zero. Other actions are less commonly mentioned. This limits what can be achieved via the ML method, because a corpus with a low share of "hits" will offer only limited information to the machine learning models and, thus, generate poor predictions.

In addition, the inter-rater agreement is not satisfactory. We hired highly competent business school undergraduate students and made great efforts to train them with guidance documents, examples, meetings and exercises. However, the rate of agreement between them was very low on many items (see the second row of Table 2). The low agreement is due to the nature of the task often being subjective. On some topics, such as whether the text discusses the firm increasing/decreasing the amount of homeworking or whether it discusses some activities to manage the risks arising from the pandemic, the tagging process is less subjective. For those the agreement is relatively high, i.e. homeworking is 0.28 and risk management is 0.30. But some topics are not as easily coded; for instance, the students found it difficult to agree on whether a text mentioned some technological innovation or whether it mentioned some improvement or deterioration in the firm's financial performance. In these cases, feedback from the students indicated that the lexicon used to describe such issues is more varied, and this introduces a greater level of subjectivity into the manual tagging process. One could no doubt

improve the level of agreement with several rounds of additional guidance, but this is a time-consuming process. The outcomes of this process therefore indicate that the ML approach is most suited to activities that appear very frequently within the corpus and which have relatively little subjectivity in the ways that they are discussed within text reports. As shown in Table 2, none of the eight activities in which we were interested met both criteria.

Table 2. Distribution of the training dataset and Cohen Kappa ratio

	Business operations	Home-working	Product & services	Technology & technological innovation	
Percentage of relevant ones	0.02	0.01	0.03	0.01	
Cohen's Kappa	0.10	0.28	0.14	0.05	
	Health & safety	Risk management	Supply chain	Financial performance	**Average**
Percentage of relevant ones	0.07	0.02	0.01	0.05	**0.03**
Cohen's Kappa	0.08	0.30	0.00	-0.12	**0.11**

Notes: Cohen's kappa is an indicator to show inter-rater agreement. There are no rules of thumb in interpretating the values. However, in general, values < 0 as indicates no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement

5.2 Dictionary-based methods

Due to the two hurdles in applying the machine learning methods, we decide to use an alternative approach. Dictionary-based methods can be effective when looking for rare events because the effectiveness of the method does not rely on the overall incidence of the activity within the document corpus, as is the case when seeking to train a ML algorithm. Here, one develops a dictionary of all words that may potentially be used to describe the action. One then searches for documents that uses one or more of these keywords. Under this method, the more comprehensive the dictionary, the better will be the resulting classification method. We use this method to identify documents relating to homeworking - the action which the students found easiest to code in the previous section. We also use it to identify documents relating to staff redundancies and hiring – these are two actions which are of interest but which were too rarely mentioned to feature in the N-gram approach.

The dictionary initially contains basic keywords from general knowledge of the language. It is then enriched with synonyms from reading the texts and looking up public resources as listed

in Table 3. The dictionary is then translated into a regular expression (regex) used in programming[7] . The regex is then tested on a manually-labelled sample of records.

The last step is iterative and involves us inspecting manually whether sentences that are tagged by the dictionary-based regex are indeed referring to the actions investigated. This process leads to refinement of the regex until it is found to achieve a certain rate of accuracy in tagging the manually-labelled records (we use a target rate of 80%).

Much of the effort is to eliminate usages of some keywords in irrelevant contexts. For example, when "redundant" is used, it should mean making *people* redundant rather than redundant *equipment* or *facilities*. We use what we call a qualifier, i.e. a bag of words that are used to refer to *people* such as worker, staff, employee, workforce, work force, etc. Appendix 7 provides an overview of the evolution of the dictionary for the action redundancy.

Table 3. Sources used for the development of dictionaries

For general language describing the action in question:
Wikipedia (e.g. https://en.wikipedia.org/wiki/Layoff)
https://www.peoplemanagement.co.uk/voices/comment/coronavirus-live-blog
https://dictionary.cambridge.org/dictionary/english/
To identify synonyms:
WordNet: http://wordnetweb.princeton.edu/perl/webwn
VerbNet: https://uvi.colorado.edu/uvi_search

The performance of the regex-based approach outside of the "regex development sample" is summarized in Table 4. We take two approaches. First, having refined the regex approach using a first set of manually-tagged records, we then evaluate the performance of the approach 'out of sample' using a second set of manually-tagged records. Second, we compare the results of the regex-based approach with the reports arising from the survey. The regex gives results at document level, but the survey is firm-level. To make meaningful comparisons, we aggregate the document tags to firm levels. Specifically, we allocate a positive tag one to a firm if any document before 24 March 2021, that is, the end time of the survey, has been tagged a positive one for this action.

We can observe from Table 4 that the regex-based approach is most accurate in terms of home-working. Comparing with the manually-tagged validation set, the regex-based method accurately classifies 63%

[7] See for an introduction to regex, for example, https://www.regular-expressions.info/quickstart.html.

of all records as to whether they mention a firm action in respect of homeworking. Some 92% of records classified via the regex-based approach as indicating a firm action in relation to homeworking are found to be 'true positives' (i.e. they do in fact relate to a document that has been manually tagged as describing a firm action in respect of homeworking) and the approach detects 64% of all records that are manually tagged as describing a firm action in respect of home working. The approach is somewhat less effective at identifying instances of redundancy, and least effective at identifying instances of hiring. We note that the lexicon around hiring is more diverse than it is for the other two actions, which may explain the lower performance for this action.

Table 4. Performance matrices of regex-term-based methods

	Home-working	Redundancy	Hiring
Using the validation document set			
Accuracy rate	0.63	0.42	0.20
Precision rate	0.92	0.35	0.27
Recall rate	0.64	0.32	0.08
Using the survey			
Accuracy rate	0.62	0.51	0.58
Precision rate	0.92	0.35	0.28
Recall rate	0.64	0.32	0.08

Notes: Accuracy rate = (TP+TN)/(TP+TN+FP+FN). Precision rate = TP/(TP+FP); The precision helps us to visualize the reliability of the model in classifying the model as positive. Recall rate = TP/(TP+FN); The recall measures the model's ability to detect positive samples. TP – true positive; TN – true negative; FP – false positive; FN – false negative.

The lower part of Table 4 describes the performance of the regex approach by validating it against the survey responses. However, we can go further by identifying the types of firm for which agreement between the regex classification and the survey response was more or less likey. This may point towards possible publication bias on the part of firms in terms of what they report on via their websites or reports.

Listed firms have the obligation to disclose information on their business activities. Therefore, we expect that for listed firms, web texts are more informative and consequently more powerful to predict actions than for unlisted firms. We investigate this issue by defining a binary (0,1) variable which identifies whether the regex classification matches the survey response (as shown by the accuracy rate in the lower half of Table 4). We then regress this binary variable on a set of firm characteristics, including its listed status, size and industry sector. Results are shown in Table 5. Few characteristics are statistically significant. For instance, larger firms appear no more likely to generate a text-based result that matches the survey than smaller firms. However, in respect of homeworking, we find that listed firms are more likely than non-listed

firms to generate an accurate prediction from the text corpus. The number of documents is also positively related to the accuracy of the prediction. This suggests that there may be more publication bias in web text around homeworking than in respect of other actions, such as redundancy or hiring.

Table 5 Validity of methods against survey: regressions

	Homeworking	Redundancy	Hiring
Listed	0.364***	-0.038	-0.070
	(0.066)	(0.074)	(0.071)
Log(turnover)	0.018	-0.016	-0.022
	(0.024)	(0.026)	(0.025)
Total number of documents (1,000)	0.036**	-0.010	-0.015
	(0.012)	(0.014)	(0.013)
Employment size dummies			
50 - 250	Reference	Reference	Reference
251-500	0.025	-0.006	-0.007
	(0.083)	(0.094)	(0.091)
501-1000	0.120	0.038	0.113
	(0.097)	(0.110)	(0.106)
1001-2000	0.015	0.183	-0.137
	(0.112)	(0.127)	(0.122)
2001-5000	0.114	0.042	0.031
	(0.110)	(0.122)	(0.119)
5000+	-0.036	-0.006	0.146
	(0.140)	(0.156)	(0.152)
Industry dummies			
A-F: Agriculture to Construction	Reference	Reference	Reference
G-I: Wholesale and retail; Transport and storage; Accommodation and food service	0.041	-0.003	-0.049
	(0.079)	(0.090)	(0.087)
J-N: Information and communication to Admin and support services	0.117	-0.051	-0.000
	(0.069)	(0.078)	(0.074)
O-S: Public admin to Other services	0.094	0.070	0.052
	(0.090)	(0.103)	(0.098)
Number of observations	310	296	308
R-squared	0.150	0.027	0.040

Notes: OLS estimations. Dependent variable = 1 if regex classification matches survey response; 0 otherwise. Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. Firms' reactions to COVID

Having explored the validity of different text classification methods, and determined that the regex-based approach is more practical and more effective than the ML approach, we now go on to use the regex-based results to explore firms' reactions to COVID in more detail. Specifically, we use the regex-based classification of homeworking, redundancy and hiring as dependent variables in regressions to identify the types of firms that undertook certain actions.

Table 6 presents the results of the regressions. Listed firms are more likely to take homeworking arrangements, make redundancies than unlisted firms. The more revenues a firm generates, the more likely a firm takes homeworking arrangements, hire and fire people. Employment size is not a strong indicator for homeworking and hiring, but it is strongly correlated with redundancy decisions. The bigger the employment size, the more likely a firm makes redundancies. We also find significant industry variations. Information and communication, administrative and support services and public administration services are significantly more likely to take homeworking arrangements because their jobs can be done at home. Those industries also make more hirings and firings. Note that we have included the total number of documents as a control to rule out the effect that listed and bigger firms tend to have more web texts than their unlisted and smaller counterparts, and therefore are more likely to get a positive one tag.

Table 6. Determinants of firm actions during the pandemic

	Homeworking	Redundancy	Hiring
Listed	0.341***	0.336***	0.047***
	(0.019)	(0.017)	(0.012)
Log(turnover)	0.056***	0.029***	0.021***
	(0.007)	(0.006)	(0.004)
Total number of documents (1,000)	0.026***	0.047***	0.043***
	(0.004)	(0.004)	(0.003)
Employment size dummies			
50 - 250	Reference	Reference	Reference
251-500	-0.014	0.025	-0.004
	(0.025)	(0.023)	(0.017)
501-1000	-0.017	0.072**	0.007
	(0.026)	(0.024)	(0.017)
1001-2000	-0.015	0.084**	-0.001
	(0.031)	(0.028)	(0.021)
2001-5000	-0.015	0.134***	0.044*
	(0.031)	(0.028)	(0.020)
5000+	-0.082*	0.141***	0.050
	(0.039)	(0.035)	(0.026)
Industry dummies			
A-F: Agriculture to Construction	Reference	Reference	Reference
G-I: Wholesale and retail; Transport and storage; Accommodation and food service	-0.017	-0.003	-0.012
	(0.023)	(0.021)	(0.015)
J-N: Information and communication to Admin and support services	0.153***	0.077***	0.045***
	(0.019)	(0.017)	(0.013)
O-S: Public admin to Other services	0.169***	0.075**	0.056***
	(0.025)	(0.023)	(0.017)
Number of observations	3487	3487	3487
R-squared	0.185	0.241	0.141

Notes: Dependent variable = 1 if web text indicates that firm has taken the relevant action; 0 otherwise. Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

7. Conclusions

Texts have attracted increasing attention over the past decade in economic analysis. Using fortnightly-scraped web texts for 3,487 firms in the UK, we show that texts can be of value to understand firm activities. Dictionary-based methods are more appropriate than machine learning algorithms based on our analysis. Texts are better at predicting positive actions due to publication bias, i.e. firms choose to cover positive things and avoid negative things on their website and reports. In addition, the predictive power is higher for large and listed firms. The methodology developed in this paper is especially useful during crises, such as the COVID pandemic, when timely measures of firm behaviour may be more difficult to collate via alternative means.

8. Appendix

Appendix 1: Sample selection

Table A1.1 Whole sample

Employment band (number of employees in 2018)	Unlisted/ Delisted	%	Listed	%	Total	%
51-250	768	22.01%	273	7.82%	1041	29.84%
251-500	426	12.21%	93	2.67%	519	14.88%
501-1000	431	12.35%	107	3.07%	538	15.42%
1001-2000	280	8.03%	76	2.18%	356	10.20%
2001-5000	516	14.79%	103	2.95%	619	17.74%
5000+	238	6.82%	178	5.10%	416	11.92%
Total count	2659	76.21%	830	23.79%	3489	100.00%

Table A1.2 Non-WMS sample

Employment band (number of employees in 2018)	Unlisted/ Delisted	Weight	Listed	Weight
51-250	696	32.29	271	1.01
251-500	392	10.87	92	1.00
501-1000	416	5.39	106	1.01
1001-2000	274	3.93	73	1.00
2001-5000	516	1.32	102	1.00
5000+	238	1.34	175	1.01

Table A1.3 WMS sample

Employment band (number of employees in 2018)	Unlisted/ Delisted	Weights	Listed	Weights
51-250	72	1.06	2	1.00
251-500	34	1.06	1	1.00
501-1000	15	1.07	1	1.00
1001-2000	6	1.17	3	1.00
2001-5000	0	NA	1	1.00
5000+	0	NA	3	1.00

Appendix 2: Description of data fields

The data is in tab-separated text files with the first row containing the name of the field. All the files have the same core fields as follows.

- crn - Registered Company Name from Firms List file
- url - url of page with Covid content
- date_read - date that the webpage was read
- title - page / section title (not always filled if not available or cannot be extracted)
- date_published - date appearing on web page that it was published (depending on nature of page, not always available)
- text - the text content mentioning the Covid related terms
- matches - the matched terms from the list of Covid related terms, semi-colon (;) separated

And additionally, for website:

- status - set to U if content has changed since last time it was delivered for url, otherwise blank.

Note the website content will contain each occurence recorded when an update has occurred in text captured.

And additionally, for news:

- source - domain of news source
- mentions - names of companies found in the text in crn:name format, semi-colon (;) separated

And additionally, for reports:

- order - incremental number indicating order extract read from report
- file_path - the filename of the corresponding PDF. the structure is 'Company ID _ Date Processed _ Last Part of URL'.
- ir_filter - Whether a document is associated with investor relations based off url and link text. True if 1, False if 0.

Appendix 3: Sample after excluding firms with problematic SIC codes

Employment band	Unlisted	Listed	Total
51-250	750	262	1012
251-500	415	87	502
501-1000	420	102	522
1001-2000	271	76	347
2001-5000	492	101	593
5000+	230	170	400
Total count	2,578	798	3,376

Appendix 4. What type of firms responded to the survey?

	A firm has responded to the survey
Listed	0.351***
	(0.019)
Log(turnover)	0.063***
	(0.007)
Employment size dummies	
51-250	Ref.
251-500	-0.018
	(0.025)
501-1000	-0.024
	(0.026)
1001-2000	-0.024
	(0.032)
2001-5000	-0.021
	(0.031)
5000+	-0.081*
	(0.039)
Industry dummies	
A-F: Primary industries; Manufacturing; Construction	Ref.
G-I: Wholesale and retail; Transport and storage; Accommodation and food service	-0.011
	(0.023)
J-N: Information and communication to Admin and support services	0.158***
	(0.019)
O-S: Public admin to Other services	0.181***
	(0.025)
N	3487
r2	0.176

Notes: OLS estimations. Dependent variable = 1 if a firm has responded the survey; 0 otherwise.
Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix 5: Descriptive statistics on documents collected from the web

Table A5.1 Number of documents collected, by type and date supplied to ESCoE

Date	Website	News	Reports	Total
16-Jul-20	64,081	33,507	3,977	101,565
31-Jul-20	2,291	9,205	8,836	20,332
13-Aug-20	23,906	8,919	21,179	54,004
27-Aug-20	33,363	11,391	3,104	47,858
10-Sep-20	30,373	5,518	42,124	78,015
24-Sep-20	16,785	10,125	14,758	41,668
08-Oct-20	20,401	8,190	13,441	42,032
22-Oct-20	16,405	11,700	8,487	36,592
05-Nov-20	11,051	8,470	12,414	31,935
19-Nov-20	13,187	13,805	14,036	41,028
03-Dec-20	11,544	11,733	16,284	39,561
16-Dec-20	10,561	4,862	12,956	28,379
31-Dec-20	11,373	12,762	14,426	38,561
14-Jan-21	9,176	9,341	6,233	24,750
28-Jan-21	10,028	9,594	7,499	27,121
11-Feb-21	9,800	8,914	6,935	25,649
25-Feb-21	8,461	10,403	15,612	34,476
11-Mar-21	9,759	11,867	14,809	36,435
25-Mar-21	10,170	12,588	21,357	44,115
08-Apr-21	9,749	12,839	18,521	41,109
22-Apr-21	9,044	12,635	18,402	40,081
06-May-21	8,749	13,279	14,117	36,145
20-May-21	8,773	10,980	14,168	33,921
03-Jun-21	7,866	11,667	13,925	33,458
17-Jun-21	7,829	8,783	11,652	28,264
01-Jul-21	5,549	9,217	12,227	26,993
15-Jul-21	5,959	10,860	10,352	27,171
29-Jul-21	6,526	10,510	10,142	27,178
12-Aug-21	8,565	9,231	14,183	31,979
26-Aug-21	7,211	7,760	10,008	24,979
09-Sep-21	6,615	6,814	8,469	21,898
23-Sep-21	6,400	12,858	9,792	29,050
07-Oct-21	8,017	10,839	11,179	30,035
21-Oct-21	7,291	7,256	10,687	25,234
04-Nov-21	6,135	7,089	12,464	25,688
18-Nov-21	7,272	6,598	6,886	20,756
02-Dec-21	6,699	8,713	10,677	26,089
16-Dec-21	6,822	9,888	10,612	27,322
30-Dec-21	7,288	14,127	11,454	32,869
13-Jan-22	5,460	11,964	5,283	22,707
27-Jan-22	6,368	10,464	9,457	26,289
10-Feb-22	7,017	7,939	7,905	22,861

24-Feb-22	6,022	6,620	8,783	21,425
10-Mar-22	5,927	4,090	13,324	23,341
24-Mar-22	3,724	3,041	12,131	18,896
07-Apr-22	3,952	3,566	11,629	19,147
21-Apr-22	2,316	2,576	12,564	17,456
05-May-22	2,243	2,003	8,394	12,640
19-May-22	2,179	1,877	9,184	13,240
02-Jun-22	2,317	1,537	9,762	13,616
Total	**518,599**	**470,514**	**606,800**	**1,595,913**

Note: Documents listed for dates in July and August 2020 were all supplied on 27th August 2020, but have been sub-divided here based on "date read"

Table A5.2 Number of documents collected, by type and month read by glass.ai

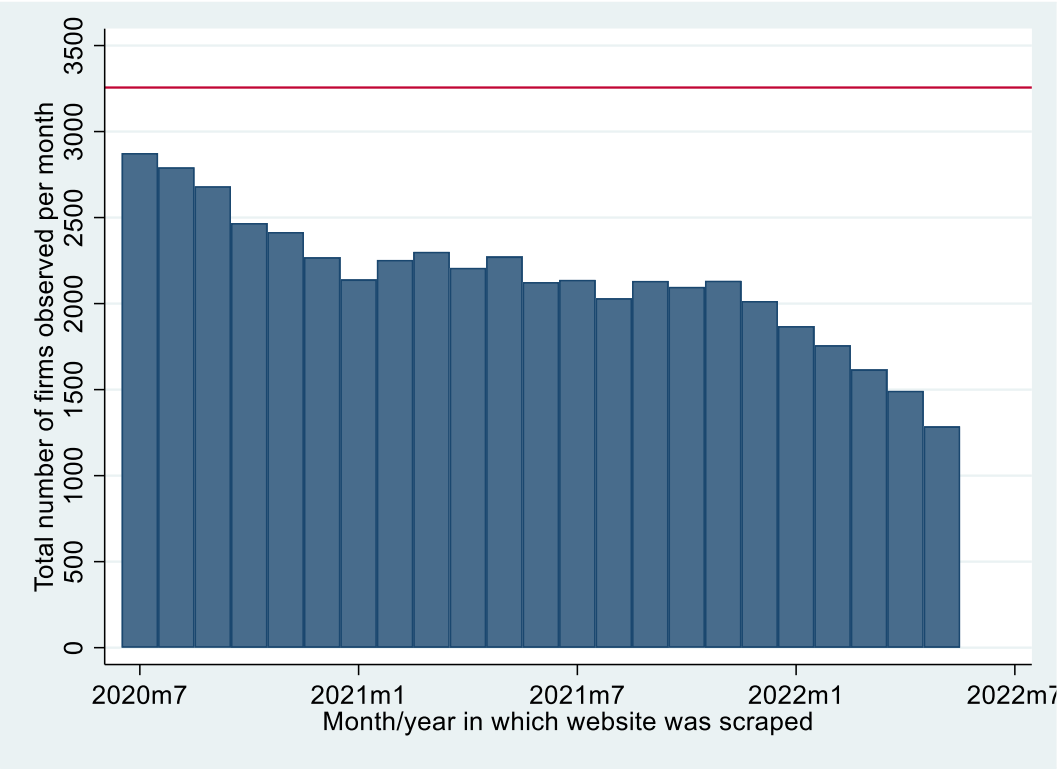
Month	Website	News	Reports	Total
2020m7	66,372	42,712	12,813	121,897
2020m8	67,253	22,078	24,283	113,614
2020m9	48,608	18,220	56,882	123,710
2020m10	30,875	21,380	21,928	74,183
2020m11	30,279	23,245	26,450	79,974
2020m12	21,934	22,552	43,666	88,152
2021m1	19,204	18,935	13,732	51,871
2021m2	18,261	23,890	22,547	64,698
2021m3	23,979	27,477	36,166	87,622
2021m4	19,002	28,234	36,923	84,159
2021m5	21,129	25,005	28,285	74,419
2021m6	13,378	18,566	25,577	57,521
2021m7	12,485	24,584	32,721	69,790
2021m8	18,275	15,311	24,191	57,777
2021m9	12,997	23,824	18,261	55,082
2021m10	15,152	15,704	21,866	52,722
2021m11	17,781	19,105	19,350	56,236
2021m12	14,110	24,015	32,743	70,868
2022m1	13,121	26,765	14,740	54,626
2022m2	12,575	11,402	16,688	40,665
2022m3	11,020	7,656	25,455	44,131
2022m4	6,289	6,035	24,193	36,517
2022m5	4,520	3,819	27,340	35,679
Total	**518,599**	**470,514**	**606,800**	**1,595,913**

Table A5.3 Number of documents collected, by month published

Month	Website	News	Report	Total
2020m1	771	316	39	1,126
2020m2	904	1,068	1,003	2,975
2020m3	8,500	6,348	9,498	24,346
2020m4	9,074	6,814	20,745	36,633
2020m5	8,776	7,272	25,933	41,981
2020m6	10,428	7,767	27,055	45,250
2020m7	9,864	16,682	32,170	58,716
2020m8	5,184	18,704	28,309	52,197
2020m9	3,594	17,572	29,751	50,917
2020m10	3,787	22,111	24,476	50,374
2020m11	3,826	26,787	38,338	68,951
2020m12	3,587	19,874	29,160	52,621
2021m1	3,483	19,571	13,863	36,917
2021m2	2,701	22,144	32,350	57,195
2021m3	3,349	27,783	67,659	98,791
2021m4	2,299	28,038	31,136	61,473
2021m5	2,201	23,785	22,996	48,982
2021m6	2,167	19,920	20,360	42,447
2021m7	2,342	24,696	22,155	49,193
2021m8	1,798	19,803	15,528	37,129
2021m9	1,901	22,824	15,618	40,343
2021m10	1,790	14,993	11,814	28,597
2021m11	1,874	18,070	16,678	36,622
2021m12	2,021	29,822	11,313	43,156
2022m1	1,914	20,327	7,248	29,489
2022m2	1,414	10,369	11,698	23,481
2022m3	1,060	6,774	21,914	29,748
2022m4	548	4,964	9,752	15,264
2022m5	326	3,530	3,500	7,356
Missing	417,116	1,786	4,741	423,643
Total	**518,599**	**470,514**	**606,800**	**1,595,913**

Note: The share of documents with a publication date is: 20% for websites; 99% for news; and 99% for reports (73% overall).

Figure A5.1 Total number of firms observed, by month read



Note: As of 2nd June 2022, 233 firms are not observed in any period. The red line sits at 3,256 (3,489-233=3,256).

Table A5.4 Share of those firms observed with specific types of document, by month read

Month	Any website	Any news	Any report
2020m7	0.91	0.51	0.14
2020m8	0.92	0.41	0.19
2020m9	0.86	0.39	0.46
2020m10	0.88	0.41	0.21
2020m11	0.83	0.44	0.25
2020m12	0.79	0.45	0.33
2021m1	0.81	0.45	0.16
2021m2	0.77	0.48	0.22
2021m3	0.80	0.49	0.26
2021m4	0.78	0.47	0.27
2021m5	0.80	0.46	0.26
2021m6	0.74	0.50	0.26
2021m7	0.70	0.51	0.31
2021m8	0.77	0.44	0.27
2021m9	0.69	0.56	0.23
2021m10	0.70	0.52	0.29
2021m11	0.73	0.52	0.25
2021m12	0.71	0.49	0.31
2022m1	0.72	0.47	0.20
2022m2	0.78	0.34	0.26
2022m3	0.79	0.29	0.30
2022m4	0.74	0.27	0.37
2022m5	0.68	0.27	0.45
Total	0.79	0.45	0.26

Note: each row is computed on the sample firms observed in that month (see Figure 1)

Table A5. 5 Number of documents per firm

	Websites	News	Reports	Total
p10	1	0	0	9
p25	12	0	0	38
p50	47	5	10	133
p75	137	40	123	411
p90	353	197	480	1,044

Note: computed on the sample of 3,256 firms with one or more documents as of 2nd June 2022

Appendix 6: Bigrams to identify commonly-occurring business activities

Topic	Bigrams
Business continuity	busi continu; busi oper; continu deliv; continu oper; continu provid; continu support; continu work; adjust oper; continu improv; remain open
Home-working	home work; remot work; return work; work home; work remot; flexibl work; hybrid work
Products & services	support custom; servic provid; product servic; new product; custom servic; busi model; product launch; product develop; product line; deliveri servic; deliveri servic; new servic; onlin retail; onlin shop
Risk management	uncertainti chang; risk uncertainti; risk manag; risk ass; reduc risk; manag uncertainti; manag risk; risk factor; risk prepar; mitig risk; minimis risk; risk analysi; econom uncertainti; extern uncertainti; high uncertainti; huge uncertainti
Supply chain	cancel suppli; chain disrupt; chain issu; chain stock; custom supplier; demand suppli; disrupt delay; disrupt suppli; global suppli; raw materi; shortag materi; suppli avail; suppli chain; suppli demand; supplier distributor; supplier manufactur
Technology & technological innovation	digit transform; new technolog; grocer technolog; technolog advanc; perform technolog; market technolog; technolog factor; digit technolog
Financial performance	actual result; balanc book; balanc sheet; cancel dividend; capit expenditur; capit return; cash flow; ~~challeng time~~; dividend cut; dividend incom; financi perform; financi posit; financi prudenc; financi report; financi result; financi statement; gross margin; growth opportun; ~~impact busi~~; incom dividend; interim dividend; market share; mover ftse; net debt; number dividend; oper profit; per share; pretax profit; profit tax; retail sale; revenu growth; share price; strong perform; year result; oper loss
Health & safety	delta variant; distanc measur; distanc workplac; emerg variant; face cover; face face; face mask; famili safe; februari pandem; fulli vaccin; govern advic; govern guidanc; govern guidelin; hand sanitis; health care; health crisi; health england; health safeti; health screen; health social; health wellb; impact infecti; indian variant; infecti respiratori; later flow; lockdown measur; lockdown restrict; mental health; nation lockdown; ongo pandem; outbreak infecti; outbreak worsen; person protect; prolong quarantin; protect equip; public health; quarantin cancel; respiratori ill; restrict ea; restrict lift; safeti measur; say virus; second wave; self isol; social distanc; spread delta; spread outbreak; spread variant; spread virus; stay home; test posit; test result; test trace; travel hospit; travel restrict; vaccin programm; wash hand; wear face; wear mask; worker care; worker test

Appendix 7: Dictionary development example

As an example, we show below the evolution of the dictionary for the action redundancy. Note that we only present root words here. Other forms including plurals and verbs in tenses other than present tense are included in the regular expressions used in our codes but excluded here for simplicity.

i. Basic keywords

{redundancy, redundant, layoff}

ii. Learning from texts

{redundancy, redundant, layoff, place/put?jobs at risk, cut?jobs}

iii. Adding synonyms

synonyms for redundancy: {job cut, job loss}; synonyms for cut: {shed, axe, reduce, lay off, downsize, sack}

iv. Translating into regex terms and refining by manual inspection

{redund?, $jobs at risk, $cut?$workers, $workers?$cut}

? represents any characters.

$ represents a set of keywords that are synonyms.

$cut includes {cut, layoff, lose, axe, shed, downsize, dismiss, sack}

$jobs includes {job, post, position, role}

$workers includes {worker, job, employee, post, position, role, people, staff, workforce}

9. References

Bellmann, L., Gleiser, P., Hensgen, S., Kagerl, C., Leber, U., Roth, D., Umkehrer, M. and Stegmaier, J., (2022) "Establishments in the Covid-19-Crisis (BeCovid): A High-Frequency Establishment Survey to Monitor the Impact of the Covid-19 Pandemic". *Jahrbücher für Nationalökonomie und Statistik*, *242*(3), pp.421-431.

Cheema-Fox, A., B. R. LaPerla, G. Serafeim and H. S. Wang (2020). 'Corporate resilience and response during covid-19', *Harvard Business School Accounting & Management Unit Working Paper*(**20-108**).

Chiripanhura B (2021) *Coronavirus and redundancies in the UK labour market: September to November 2020*, London: Office for National Statistics.

Dorr et al (2022) An Integrated data framework for policy guidance during the coronavirus pandemic (PLoSOne, 17, 2)

García, R., Gayer, C., Hölzl, W., Payo, S., Reuter, A. and Wohlrabe, K. (2020) *The impact of the COVID-19 crisis on European businesses: Evidence from surveys in Austria, Germany and Spain* (No. 31). EconPol Policy Brief.

Gough J (2020) *Business insights and impacts on the UK: 19 November 2020*, London: Office for National Statistics.

Hassan, T. A., S. Hollander, L. Van Lent, M. Schwedeler and A. Tahoun (2020) *Firm-level exposure to epidemic diseases: Covid-19, sars, and h1n1*, NBER WP 26971. National Bureau of Economic Research.

HM Revenue and Customs (2021) *Coronavirus Job Retention Scheme Statistics: 16 December 2021*, London: HM Revenue and Customs.

HM Treasury (2021) *HM Treasury Coronavirus (COVID-19) Business Loan Scheme Statistics*, 8th July 2021, London: HM Treasury.

Hopson E (2020) *Business Impact of Coronavirus, Analysis Over Time, UK: Waves 2 to 5 Panel*, London: Office for National Statistics.

Kinne, J., M. Krüger, D. Lenz, G. Licht and P. Winker (2020). 'Coronavirus pandemic affects companies differently', in (Editor Ed.)^Eds.), *Book Coronavirus pandemic affects companies differently*, City: ZEW–Leibniz Centre for European Economic Research Mannheim.

Mizen P, Analyi L, Bloom N, Thwaites G and Yotzov I (2022) "Two years on, how has the pandemic affected businesses in the UK?", Economics Observatory blog, retrieved 7th July 2022.

Sharma et al (2020) COVID-19's impact on supply chain decisions: strategic insights from NASDAQ 100 firms using Twitter data (Journal of Business Research, 117)

Valero A, Riom C and Oliveira-Cuhna J (2021) *The Business Response to Covid-19 One Year On: Findings from the Second Wave of the CEPCBI Survey of Technology Adoption*, London: Centre for Economic Performance.