Investigating Businesses' Experience of COVID via Web Crawling and Text Mining

John Forth, City University of London
Charlotte Meng, Queen Mary
Rebecca Riley, King's College London

ESCoE Annual Conference
King's College London, 17 May 2023

Outline

- Background
- Motivation
- Main findings
- Data collection
- Results
- Conclusions

Background

- Business data are conventionally collected from companies' administrative accounts (e.g. tax records, trade data) and surveys (e.g. Annual Business Survey).
- Administrative account data cannot cover all aspects in the economy and lag behind markets.
- Surveys are expensive to run, and low responses rates can be an issue, especially during crises.

Motivations & research approach

- We aim to gather information on UK business activities from public information sources, using the COVID pandemic as a case study.
- Public information is fast, diverse, and usually in large volumes.
- We collect fortnightly data for approx. 3,500 firms, searching their corporate websites and online news sources for text relating to the pandemic
- We use text analysis methods (e.g. topic modelling) to interpret the data.

Research questions

- **Discovery**
  - How did COVID affect businesses' operations?
- **Audit**
  - To what extent can public information provide reliable insights into businesses' responses to COVID (complementing – or substituting for – surveys)?
- **Analysis**
  - Can such public information be used to explain heterogeneity in firm performance?
  - *Coming later in the project…*

Data collection: Sample selection from FAME

- In June 2020, we drew a stratified random sample of 4,135 firms with 51+ employees in SIC(2007) Sections A-S
  - Larger firms and listed firms over-sampled
  - Representative of 61% of private sector employment & 70% of private sector output
- Financial performance data available for later analysis
- Firms without a website (11%) or sharing a website with another firm in the sample (4%) were excluded, leaving 3,489 firms

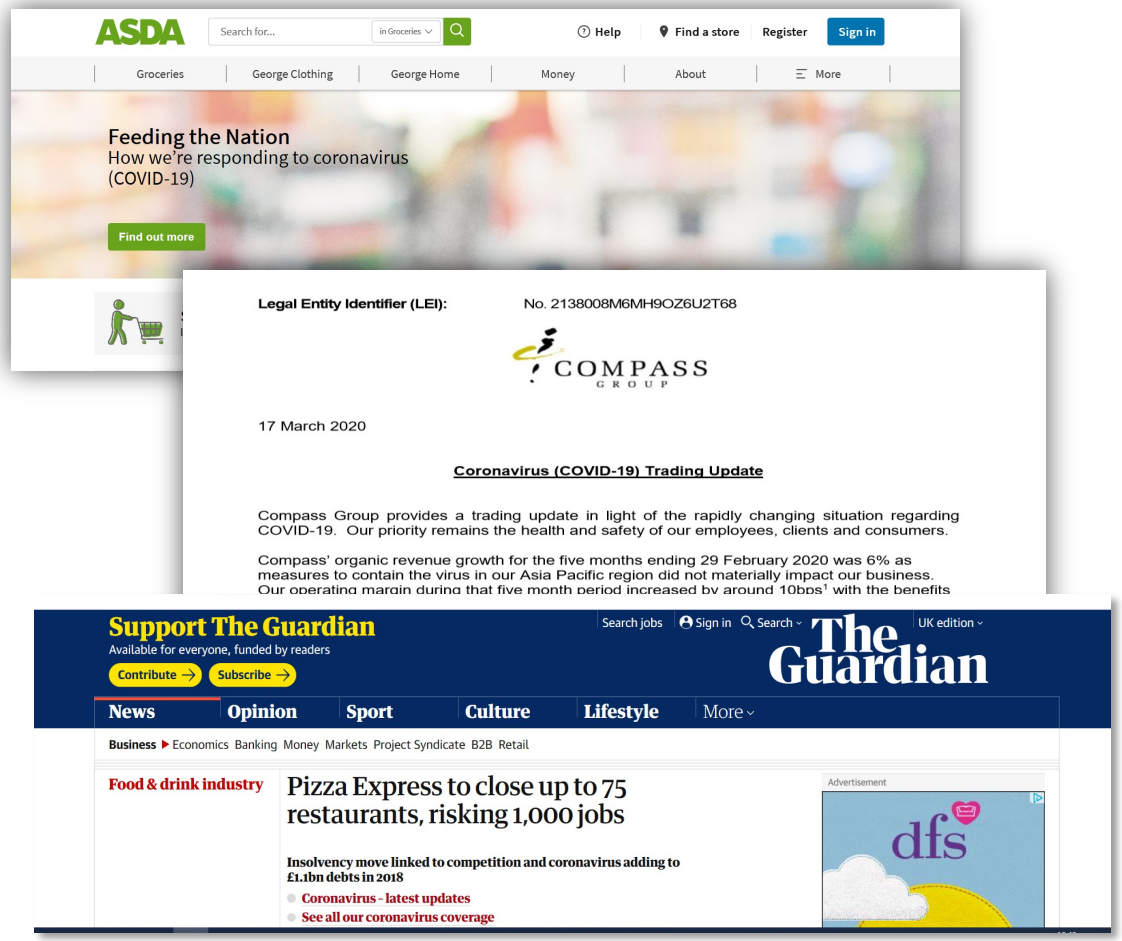
Data Collection

We collect COVID-related content on these 3,489 firms fortnightly from 8th July 2020 – 15th July 2022 (~50 waves).

- Company websites – broad coverage of companies but potentially selective on issues reported
  - HTML content
  - Press releases
  - IR micro-sites

- Online news outlets (e.g. Guardian, BBC, FT, local news) – high-profile companies but all issues of public interest

We also conducted a traditional survey of these firms in March 2021, with a 9% response rate (n=311)



Extracting COVID-related content from the web

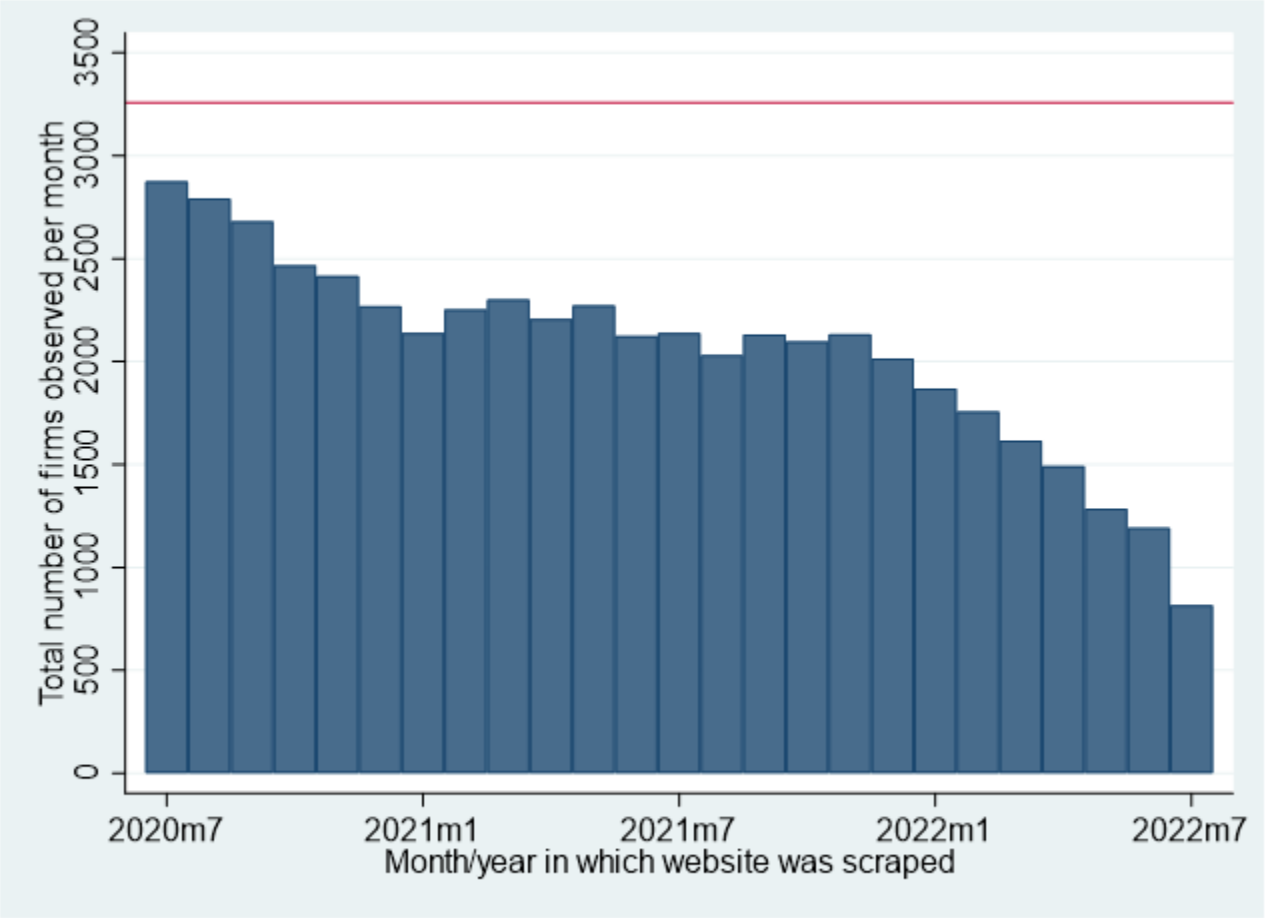
- We identify any document (webpage, PDF, news report) containing any one of a set of COVID-related keywords
  - Coronavirus / “corona virus”/corona-virus
  - Covid / covid19 / covid-19
  - SARS-Cov-2
  - Omicron (added 17th Dec 2021)
- We then extract 100 words before and after the COVID keyword -> the corpus of text documents

Data collected

- 2.1m documents mentioning a COVID keyword
  - 1.42m documents from company website
  - 0.64m documents from news sources

crn	url	date_read	date_publ	title	text	matches
10758801	https://www.energean.com/s	09/07/2020		Build a better future	We have called on the PM to create a green recovery that is just for	covid;covid-19
10758801	https://www.energean.com/s	09/07/2020		We care and we figh	The fight against COVID-19 is continuous! Energean has already und	covid;covid-19;pandemic
01610897	http://www.adareinternationa	29/07/2020		Launching our new S	To see our full range of PPE, safety signage and other essentials you	lockdown;coronavirus;ppe
01610897	http://www.adareinternationa	29/07/2020		Supporting councils	The UK government has launched many initiatives since the COVID-	reopening;covid-19
01610897	http://www.adareinternationa	29/07/2020		How are we reacting	In light of the evolving situation regarding the spread of Coronaviru	coronavirus;covid-19
07033534	http://www.novatech.co.uk/b	08/07/2020	22/06/2020	5 Essential Technolo	The office of the future promises to bring with it a new approach wl	lockdown;social distancing;coronavirus;pandemic
01131910	http://www.dcnorris.com/nev	13/07/2020		Food Processing Tec	Learn about our response to Coronavirus (COVID-19). Find Out More	coronavirus;covid-19
00775598	https://www.oxinst.com/new	13/07/2020	02/07/2020	Block Listing Six Mor	COVID-19 update The interests and wellbeing of our employees an	covid-19
00775598	https://www.oxinst.com/new	09/07/2020	08/04/2020	Expanding Access to	Expanding Access to our Satellite Licenses Greater flexibility during	sars-cov-2
00596137	http://www.gpe.co.uk/news-r	13/07/2020	20/05/2020	Creation of GPE Cov	The Fund will be open to all GPE employees to contribute to and th	covid-19
00596137	http://www.gpe.co.uk/news-r	13/07/2020	20/05/2020	Annual results 2020:	I am pleased to report on a strong operational performance for the	pandemic;covid
02714645	https://www.zotefoams.com/	09/07/2020	09/04/2020	Plastazote® polyeth	Medical device manufacturer J-Pac Medical is now producing face sl	ppe
02715398	https://www.ingevity.com/ne	09/07/2020	02/06/2020	Ingevity to reduce c	Jun 02, 2020 Ingevity to reduce costs, staffing in response to corona	furlough;coronavirus;pandemic;covid-19
06454113	https://www.willmottdixon.c	29/07/2020	30/06/2020	Contractors Declare	Willmott Dixon invites others to commit to limiting global warming	covid-19

Number of firms mentioning Covid in each month read



Note: 231 firms are not observed in any period.

Results: How did COVID affect businesses' operations?

Topic modelling with LDA (Latent Dirichlet allocation)

1. Data pre-processing: tokenization, remove stop words, remove punctuations and symbols, stemming, etc.
2. Model training
3. Interpret topics and examine the distribution of topics

Source: Adapted from Hannigan et al. 2019 (AMA)

Example topics and keywords

Topic 7: work, office, employee, home, remote, business, team, new, client, working → **Homeworking**

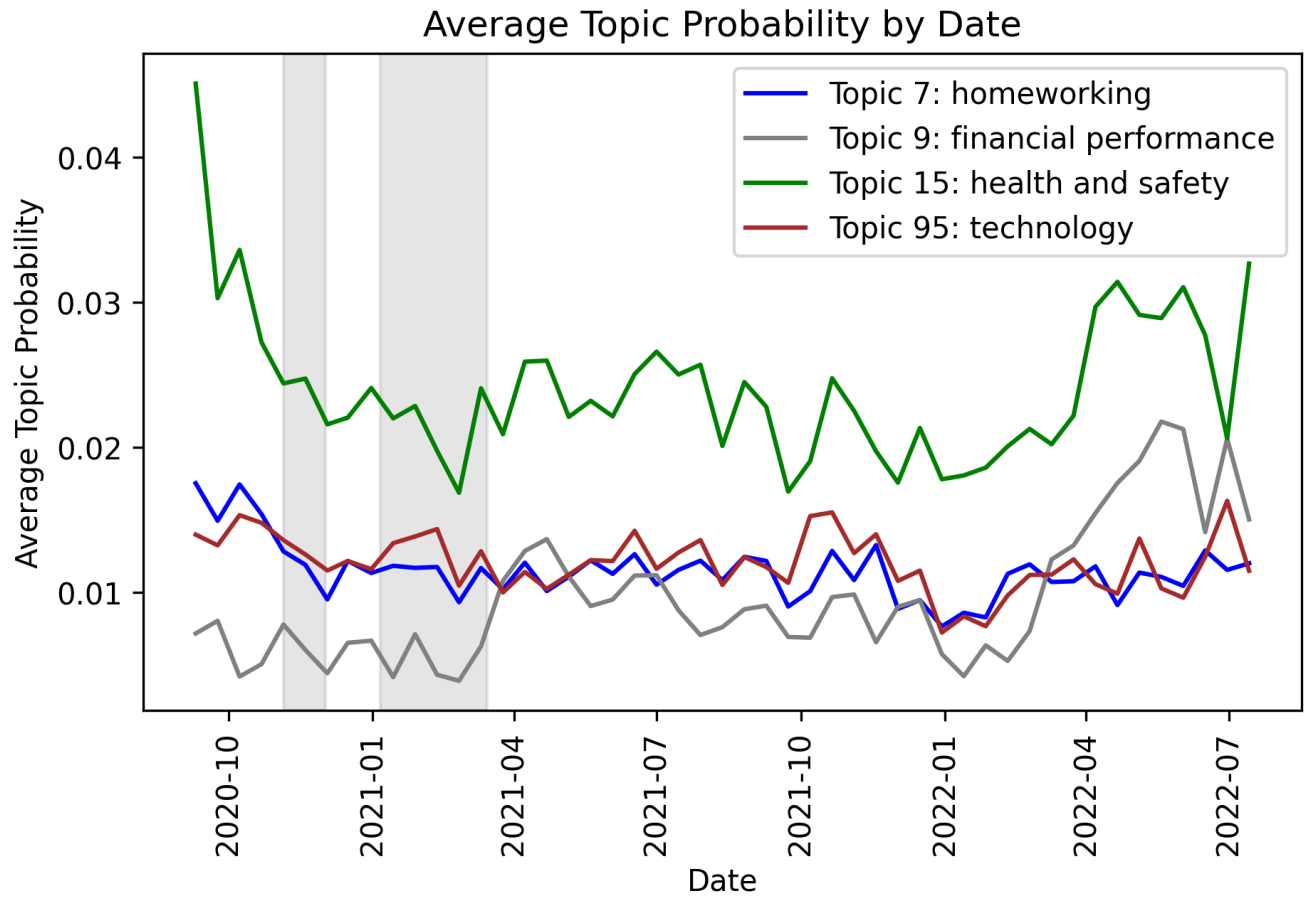
Topic 9: financial, audit, statement, report, group, company, auditor, information, annual, director → **Financial performance**

Topic 15: safety, work, health, employee, ensure, continue, site, measure, business, support → **Health and safety**

Topic 95: digital, technology, datum, service, solution, cloud, business, system, software, platform → **Technology**

Results: How did COVID affect businesses' operations?

Distribution of topics over time



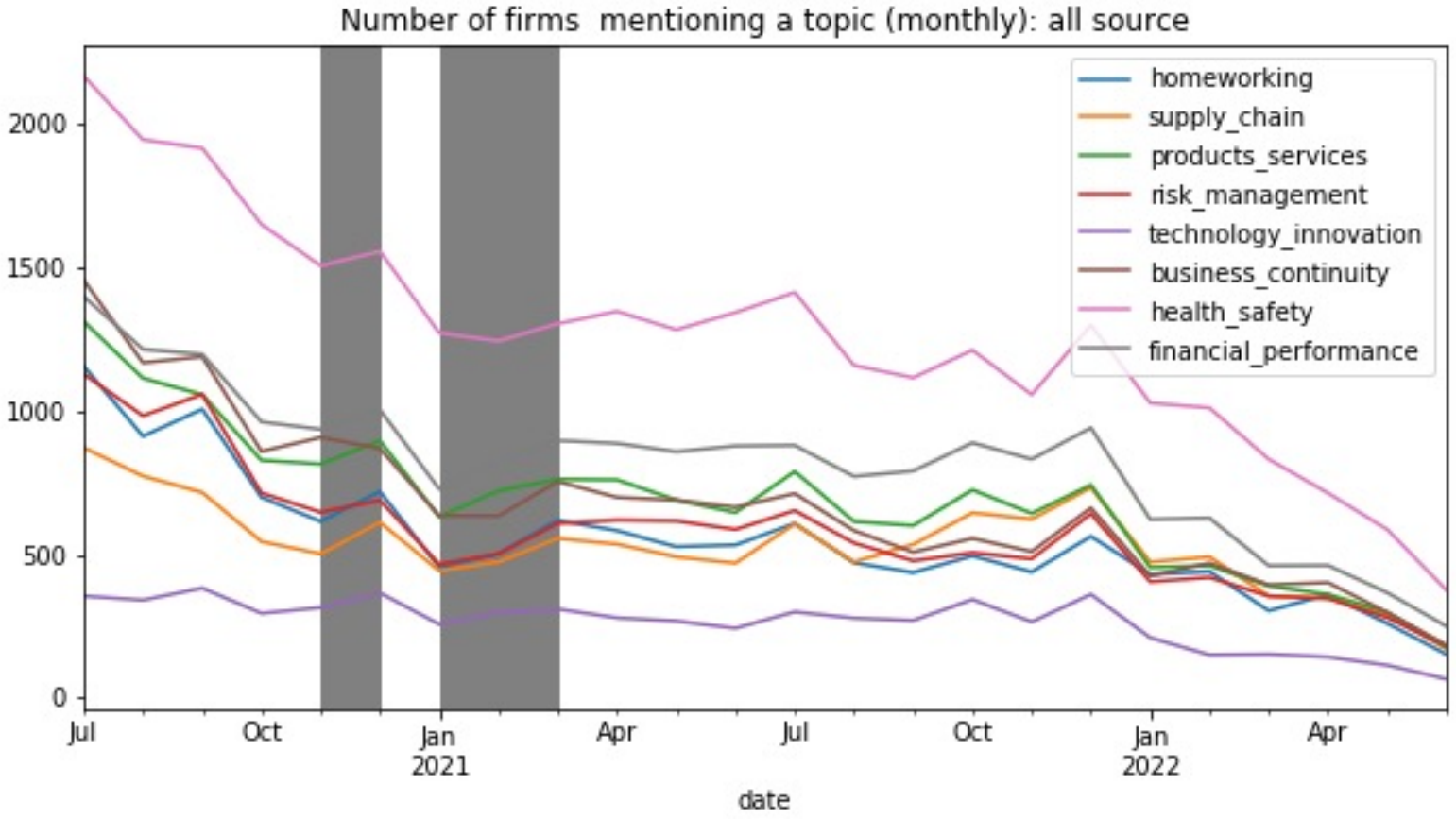
Results: N-gram

1. Identify the top 10,000 commonly-occurring unigrams and bigrams
2. Identify eight categories of business activity that are commonly mentioned
3. Go through the top 1,000 bigrams to compile a list of terms that occur under the above eight categories. Examples in the following table.
4. Tag the documents and aggregate to firm-time levels to determine if something has been published about the firm in relation to a certain business activity in a certain period.

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

Results: N-gram

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



Results: Can public information provide reliable insights into businesses' responses to COVID?

Total error framework:

- Coverage errors – are certain types of firm over or under-represented?
- Measurement errors – can we make reliable inferences from textual analysis?

Coverage

Profile of the selected sample and achieved samples

	Selected sample	Firms with valid website	Any COVID-related documents	Survey sample
	%	%	%	%
Not listed	79.8	76.2	74.3	75.2
Listed	20.2	23.8	25.7	24.8
Employment size:				
51-250	28.4	29.8	28.3	30.9
251-500	15.1	14.9	15.0	18.0
501-1000	15.0	15.4	15.6	12.5
1001-2000	10.4	10.2	10.6	9.0
2001-5000	18.9	17.7	18.1	17.7
5001+	12.1	11.9	12.5	11.9
Total	4,135	3,489	3,081	311

OLS regression analysis of document intensity: all firms

	(1) Any document (Company website)	(2) Number of documents (Company website)	(3) Any document (News reports)	(4) Number of documents (News reports)
Listed	0.034***	2.281***	0.262***	3.018***
	(0.003)	(0.154)	(0.003)	(0.161)
Log(turnover/ employment)	0.057***	1.236***	0.034***	1.112***
	(0.001)	(0.050)	(0.001)	(0.040)
Employment size (Ref. <250 employees)				
251-500	0.041***	0.844***	0.041***	0.573***
	(0.003)	(0.122)	(0.003)	(0.034)
501-1000	0.087***	1.039***	0.065***	0.834***
	(0.003)	(0.092)	(0.003)	(0.044)
1001-2000	0.109***	2.495***	0.098***	1.124***
	(0.004)	(0.239)	(0.003)	(0.053)
2001-5000	0.172***	3.184***	0.158***	4.060***
	(0.003)	(0.164)	(0.003)	(0.283)
5000+	0.263***	7.449***	0.254***	8.853***
	(0.004)	(0.323)	(0.004)	(0.379)
Industry FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
N	165032	165032	165032	165032
r2	0.115	0.020	0.176	0.019

Notes: Standard errors in parentheses.
* p < 0.05, ** p < 0.01, *** p < 0.001

Measurement

We focus on two topics covered in the n-gram approach, discussed earlier

- Use of homeworking
- Use or adoption of new technology

Challenges:

- Commentary vs reportage
  - Web content not informative for knowledge brokers (media companies, professional services firms)
- Content vs meaning
  - Has done *[action]*? Is planning to do *[action]*? Has decided not to do *[action]*?

Results: ML

ML approach doesn't work well due to:

- imbalanced training dataset
- low inter-rater agreement.

	Home-working	Technology & technological innovation
Percentage of relevant ones	0.01	0.01
Cohen's Kappa	0.28	0.05

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

Results: Dictionary method

A dictionary development example

1. Basic keywords

{technology, digital}

2. Learning from texts

{new technology, technology transform, technology advance, digit transform, digit innovation, adopt (enterprise resource planning, customer relationship management, remote working technology, cloud computing, mobile technology, automated machinery, AI, 5G, Microsoft Teams, Zoom, video conferencing, online sale, online marketing) }

3. Adding synonyms

synonyms for AI: {artificial intelligence}; synonyms for adopt: {extend, expand, invest, introduce, launch}

4. Translating into regex terms and refining by manual inspection

{new tech*, tech* transform*, digit* transform*, digit* innovat*, tech* advanc*, $adopt(.*)$spec_tech, $spec_tech(.*)$adopt}, *where*

$ adopt includes {adopt, extend, expand, invest, introduce, launch}
$ spec_tech includes {enterprise resource planning, customer relationship management, remote working technology, cloud computing, mobile technology, automated machinery, AI, artificial intelligence, 5G, Microsoft Teams, Zoom, video conferencing, online sale, online marketing}

Results: Dictionary method

Performance matrices of regex-term-based methods

	Homeworking	Technological innovation
Using manually-tagged records:		
Accuracy rate	0.63	0.87
Precision rate	0.92	0.94
Recall rate	0.64	0.87
Using the survey:		
Accuracy rate	0.62	-
Precision rate	0.92	-
Recall rate	0.64	-

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.

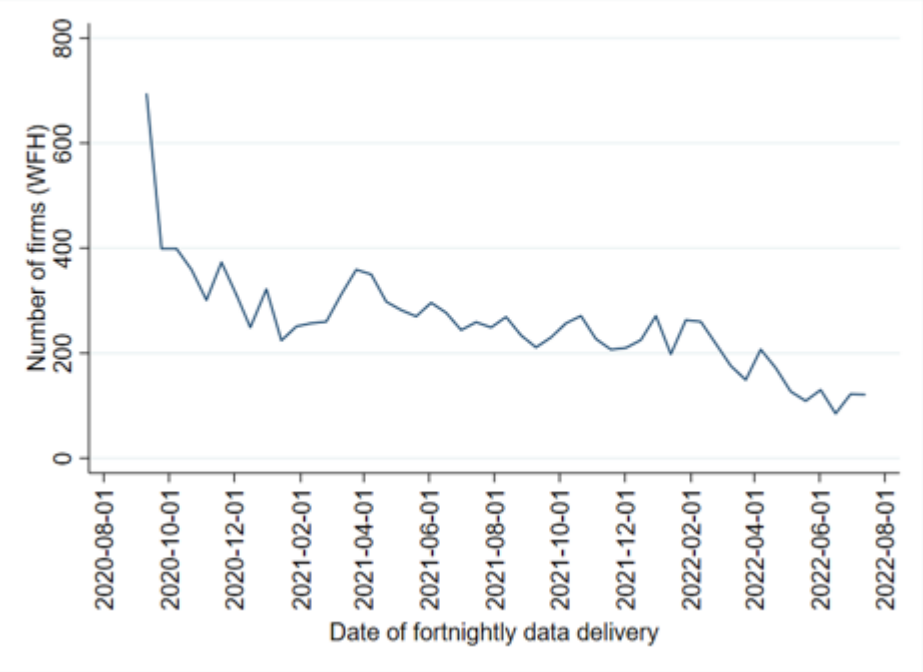
Validity of methods against survey: regressions

	Homeworking
Listed	0.364***
	(0.066)
Log(turnover)	0.018
	(0.024)
Total number of documents (1,000)	0.036**
	(0.012)
Employment size dummies	
50 – 250	Reference
251-500	0.025
	(0.083)
501-1000	0.120
	(0.097)
1001-2000	0.015
	(0.112)
2001-5000	0.114
	(0.110)
5000+	-0.036
	(0.140)
Industry dummies	
A-F: Agriculture to Construction	Reference
G-I: Wholesale and retail; Transport and storage; Accommodation and food service	0.041
	(0.079)
J-N: Information and communication to Admin and support services	0.117
	(0.069)
O-S: Public admin to Other services	0.094
	(0.090)
Number of observations	310
R-squared	0.150

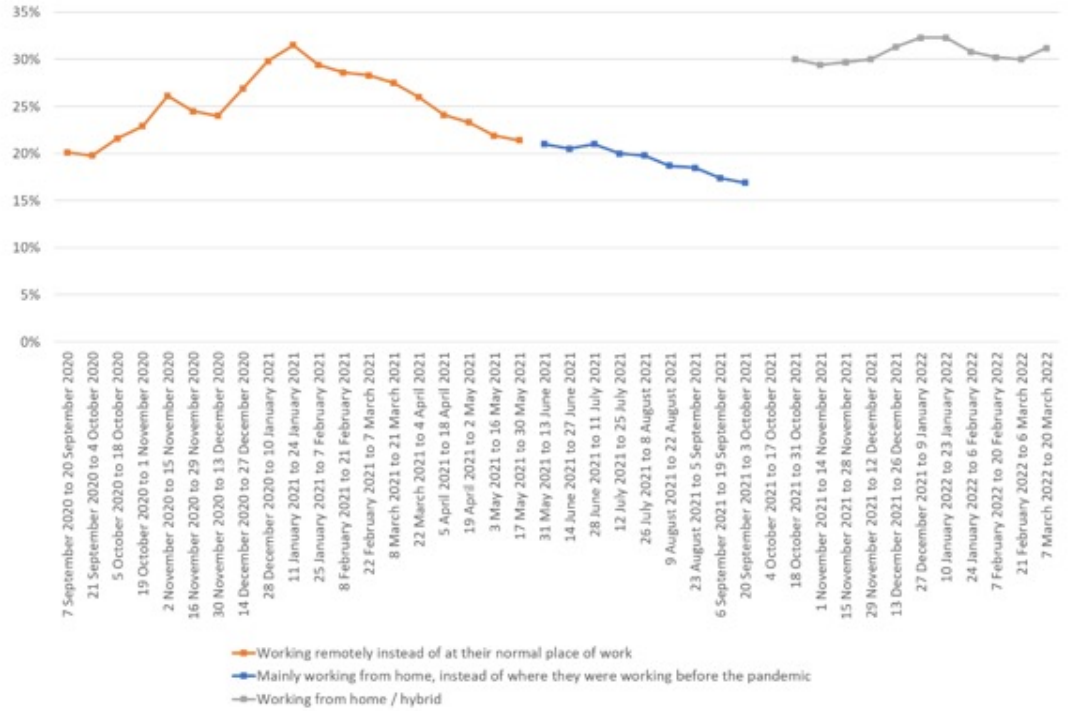
Notes: OLS estimations.
Dependent variable = 1 if regex classification matches survey response; 0 otherwise.
Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Results: Dictionary method

Discussion of homeworking in text data over time (Sept 2020-July 2022)



Percentage of employees working from home, as reported by firms (Sept 2020-March 2022)



Source: ONS Business Impact of Coronavirus Survey.

Conclusions

- We find that it is possible to build a dataset that is representative of the population, avoiding some of the non-response problems (low N) that might affect a traditional survey.
- The textual data cover a wide range of business experiences, but extracting valid measures of businesses' responses (akin to those obtained from a survey) brings a series of challenges
  - Commentary vs reportage
  - Content vs meaning
- Businesses' responses are difficult to classify via machine learning algorithms
- Dictionary-based methods are more effective, but performance varies (better for listed firms)

The results are preliminary. Comments and suggestions will be greatly appreciated.

For correspondence and updates to the paper

Charlotte Meng (QMUL)

c.meng@qmul.ac.uk