Economic impact of Covid in Scotland: pipeline and methodology plan

Nesta

October 14, 2020

Contents

1 Enumeration and description of questions and tasks

1.1 Policy questions

1.1.1 1. What do business website notices say about exposure/response?

Data used: Covid Notices posted on business websites
Tasks:

1. Aggregate by geography and sector and report distribution, identifying where there are there more/less notices than expected.
2. What proportin of notices contain information specific to the business? How many notices change between snapshots?
3. Train a topic model pipeline on the Covid notice text
4. Generate a measure of exposure based on the topic model, and topics identified as relating to business exposure
5. Train a sentiment analysis model on the Covid notice text

- **Stretch:** High risk and effort

Outputs:

1. Geographic and sectoral distribution of notices (comparing to expected levels based on business prevalence)
2. Topic trends
3. A topic-based measure of Covid exposure
4. List of businesses which have adopted new strategies
    - **Stretch:** High risk and effort

Scale-up considerations:

1. Will require generating a topic-based measure of exposure that remains relevant over time with minimal intervention.

1.1.2 2. What do social media posting say about exposure/response?

Data used: Twitter features generated by tasks 16 and 17
This question will need to be restricted to two micro-pilots (one geography and one industry). See tasks 16 & 17 for more context.
Data used: Twitter API and Glass organisation data (for seed URL's)
Tasks:

1. Generate Covid exposure measure based on Twitter features

Outputs:

1. Twitter-based measure of Covid exposure

Scale-up considerations:

1. Collection, processing, and storage costs
2. A scale-up could look at the textual content of tweets, calculating things such as their sentiment; however this will require a lot of R&D.

1.1.3 3. What are the levels of adoption of technologies/strategies such as e-commerce?

Data used: Covid Notices and Glass e-commerce flag
Tasks:

1. Analyse prevalence and changes in the glass e-commerce flag

- **Outcome:** Data quality issues make this task not viable

2. Detect a change of strategy using NLP on Covid notices

   - **Stretch goal:** High risk and effort
   - **Note:** If this task were successful, in a scale-up phase this could be applied to Twitter data

Outputs:

1. List of businesses which have adopted new strategies

1.1.4 4. Sectoral and geographic distribution of exposure (Google trends)

Data used: Glass business descriptions, Glass-CH fuzzy-matched data, NOMIS labour market statistics, and Google trends API
Tasks:

1. Aggregate company descriptions by sector (using the Glass-CH data) and extract salient terms
2. Query Google trends API for terms
3. Measure change in search popularity compared to pre-pandemic
4. Generate sector exposure indicator
5. Generate regional exposure indicator (using labour market statistics to map across from the sector exposure indicator)

Outputs:

1. Sectoral and regional exposure indicator using google trends data

1.1.5 5. What are the diversification opportunities?

Data used: Glass business descriptions, Glass-CH fuzzy-matched data, and NOMIS labour market statistics
Tasks:

1. Train model to predict (multiple) sector probabilities based on business descriptions

   - SIC codes from the Glass-CH matching are used as labels

- Business description text is used as a feature
- Businesses receive a probability of belonging to each industry

2. Construct measure of sector similarity based on frequency of sector co-occurrence - if many businesses have a high probability of belonging to two sectors then those two sectors will have a high measure of similarity.

3. Use sector similarity as a proxy measure for the diversification options of a business within a given sector

4. Create geographical measure of diversification options (using labour market statistics to map across from the sector exposure indicator)

Outputs:

1. Sectoral and geographic measure of diversification opportunities

Scale-up considerations:

1. How to establish whether the diversification pathways identified are realistic?

1.1.6 6. What is the exposure to Covid of innovative companies - STRETCH GOAL

Data used: Gateway to Research, PATSTAT, and other Covid exposure indicators from other tasks
Due to the lower priority of this item and the significant complexity added by introducing two large datasets not used elsewhere this task has been deferred and labelled as a stretch goal.
Tasks:

1. Clean, process, and fuzzy-match Gateway to Research to Glass

2. Clean, process, and fuzzy-match PATSTAT to Glass

3. Compare the effect of presence in innovation datasets (GtR and PATSTAT) on various exposure indicators generated from other tasks.

Outputs:

1. A list of 'innovative' businesses

2. The relationship between being 'innovative' and levels of Covid exposure (by geography and industry)

1.1.7 7. What is the relationship between exposure and business failure?

Data used: Covid exposure indicators, business failure data
What effect do the different Covid exposure indicators have on the levels of business failures?
Tasks:

1. Effect of high industry / place exposure (google trends) on business failure
2. Effect of a business being 'innovative' on business failures
3. Effect of the presence of e-commerce / new strategy or technology adopted
    - **Note:** These indicators have been deemed not viable and deferred respectively
4. Effect of Covid notice exposure indicator on business failure
5. Effect of Social media Covid notice exposure on business failure
6. Effect of poor diversification opportunities on business failure

Outputs:

1. Correlation between each exposure indicator and business failure

1.1.8 8. Can we identify businesses at risk of failure? Combine measures of exposure to predict failure.

Data used: Covid exposure indicators, business failure data,
Tasks:

1. Build a predictive model for business failure, combining the features and insights from Task 7.
2. Identify businesses most at risk of failure

Output:

1. List of businesses most at risk of failure
2. Predictive model

Scale-up considerations:

1. An automated pipeline that collects new data to generate exposure indicators to feed into the model will require a large effort
2. Business failure data would need to be collected continuously to detect any drift in the model's accuracy
    - Note: Depending on the type of model trained, the model could continuously update as new data arrived

1.1.9 9. Indirect impacts - NOT VIABLE

Data used: Glass network data, I/O tables
Not viable due to data quality issues with the Glass network data, and a weak rationale for this task purely based on I/O tables data.

1.1.10 10. Detect business failures in a timely manner

Data used: Companies House, TBD
There is a risk that the level of timeliness and granularity of available business failure data will limit the possibilities for downstream tasks: 7 & 8.
Tasks:

1. Identify timely data-sources of business failure, such as:
    (a) Companies House
        - There is a large lag (>3 months) in Companies House and only a subset of business failures are captured
    (b) `https://www.gov.uk/government/statistics/incorporated-companies-in-the-uk-apr`
    (c) `https://www.gov.uk/government/statistics/monthly-insolvency-statistics-june-2`
2. Collect business failure data
3. Assess timeliness and accuracy of business failure data

Output:

1. Levels of business failure at the sectoral and regional level
2. Levels of business failure at the business level (if-available)

1.2 Data issues

No steps are explicitly dependent on these points, but time invested into these will correspond to an increase in data quality.

1.2.1 11. How to deal with multi-site firms

1. How to resolve conflicting information between trading address (Companies House) and the addresses (up to 5) found by Glass AI.
2. How to discern a true multi-site firm from a false one (e.g. a second address is detected that corresponds to e.g. the address where an event is to be held)
3. What if there are more than 5 sites?

1.2.2 12. What is the coverage and bias of the business website data?

Data used: All Glass data, Companies House data, NOMIS labour market statistics
Tasks:

1. Data quality and profiling of each Glass data table
2. Exploratory data analysis
3. Compare coverage in Glass with Companies House and NOMIS labour market statistics (BRES and IDBR)

Output:

1. Report of bias and coverage

1.2.3 13. What is the most reliable way to match industrial taxonomies? Glass AI uses Linkedin's industry list, Companies House uses SIC codes, and we wish to report SIC codes.

Whilst we can get SIC codes for many Glass businesses by fuzzy-matching to Companies House, this isn't possible for every business.
The model trained in task 5 allows for generation of (multiple) sector labels which is one solution to the problem; however a mapping between the two taxonomies would provide an additional validation layer.
Given alternatives (task 5) exist, this task can probably be deferred.

1.3 Engineering "tasks"

1.3.1 14. Data modelling of Glass datasets

Data used: All Glass data
Tasks:

1. Schema
2. Data validators
3. Consistent merging of multiple snapshots

1.3.2 15. Fuzzy-matching

Data used: Glass organisation data, Companies House
Tasks:

1. Develop reusable pipeline
2. Consider how to validate accuracy of fuzzy-matching
3. Choose precision-recall trade-off

1.3.3 16. Finding social media links

Data used: Glass list of business URL's
This requires Nesta to (re)scrape websites of organisations to find links to social media handles.
Legal limitations of scraping facebook data and the lack of a suitable API limit us to considering Twitter data only.
This question will need to be restricted to two micro-pilots (one geography and one industry). Collection, processing, and storage for all data will take too much time and money for a pilot. Clackmannanshire has been proposed as a geographic target and "Accomodation and food services" as a sectoral target; however we will need to restrict such a broad sectoral target to Scotland only.
Tasks:

1. (Re)scrape business websites to search for twitter handles

Outputs:

1. List of social media accounts for glass organisations in the two micro-pilots

1.3.4 17. Fetch twitter data for organisations with social media links.

Data used: Twitter API, Business website social media links (Task 16)
As with task 16, this question will need to be restricted to two micro-pilots (one geography and one industry).
Tasks:

1. Assess Twitter API costs and rate-limits
    - Every 24 hours we can fetch ~100,000 tweets which is sufficient for a pilot stage but would be insufficient to collect historic tweets (there are likely >100,000 new tweets being generated a day across the organisations within Glass)
    - Limitation: you can retrieve the last 3,200 tweets from a user timeline, therefore if they tweet frequently then we may not be able to far back enough in time.
    - There is the potential to apply for an academic research account which would give better access for free: `https://developer.twitter.com/en/solutions/academic-research`
    - The "Enterprise" pricing of paid access to the Twitter API (which allows full historical access rather than the last 30 days or last 3,200 tweets) is unclear but costs at least $99/month
2. Collect tweets from Twitter API
3. Generate features based on the frequency of tweeting and account longevity (ignoring the textual content of tweets)

Outputs:

1. Twitter dataset capturing tweet behaviour and account longevity

2 Task graph and schedule

It is structured into 4 separate phases (explained further below). Task 8 (identify businesses at risk of failure) builds on almost every component and could therefore be seen as an 'end-goal'; however an MVP of 8 does not require all components to be complete, only some. The 4 phases are structured such that we can deliver 8 (and thus many of the components it depends upon) in an agile way. Achieving a first iteration of 8 could be seen as our MVP.

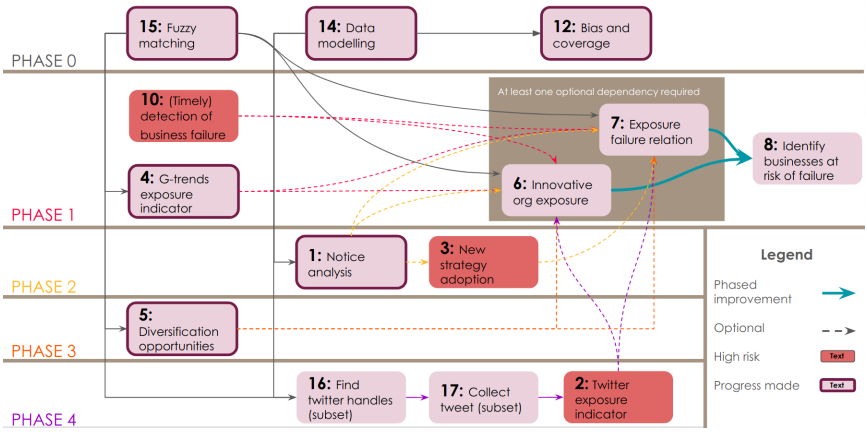

Figure 1: Task graph showing dependencies between steps, organised according to a phased delivery plan.

The order that phases 2, 3, and 4 are performed in is interchangeable; however performing phase 4 last is preferable.
Should 6, 7, and 8 be de-prioritised by SG or should we be unable to obtain business failure data of sufficient quality then the phases still work; however more time could be invested in other tasks.

2.1 Phase 0 - Groundwork

This phase lays the groundwork for the rest of the project by providing exploratory analysis and the fuzzy-matching tool that is used in many tasks.

2.2 Phase 1 - MVP

This phase achieves a MVP of tasks 6, 7, and ultimately 8.

2.3 Phase 2 - Iteration

6, 7, and 8 are enhanced by adding new indicators of exposure/adoption.

2.4 Phase 3 - Iteration

6, 7, and 8 are enhanced by adding indicators of exposure based on the diversification opportunities available to a company.

2.5 Phase 4 - Iteration

6, 7, and 8 are enhanced with an exposure indicator derived from tweet frequency.