Pipeline plan

Alex Bishop

October 7, 2020

Contents

1 Pipeline planning

1.1 Policy questions

1. What do business website notices say about exposure/response?
   - (a) Topic evolution over: time, space, industry
   - (b) Where are there more or less notices than expected?
   - (c) Sentiment analysis [HARD]
2. What do social media posting say about exposure/response? Perform for 2 subsets of data: one region and one industry
   - (a) How often are social media posts updated?
   - (b) Longevity - when do accounts start and stop posting?
   - (c) Sentiment analysis [HARD]

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

    (a) Are there changes in the glass e-commerce flag [DQ ISSUES]

    (b) Detect a change of strategy using NLP on notices or social media activity [HARD]

4. Sectoral and geographic distribution of exposure

    (a) Aggregate company descriptions by sector and extract salient terms

    (b) Query Google trends for terms

    (c) Measure change compared to pre-pandemic

    (d) Generate sector exposure indicator

    (e) Generate regional exposure indicator

5. What are the diversification opportunities?

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

    (b) Construct measure of sector similarity based on frequency of sector co-occurrence

    (c) Use sector co-occurence as a proxy for diversification options

    (d) Construct "sector-space"

    (e) Aggregate geographically to create geographical measure of diversification options

6. What is the exposure to Covid of innovative companies

    (a) Compare the effect of presence in datasets (GtR and PATSTAT) on various exposure indicators.

    Note: could be skipped if not deemed important by SG - it adds significant complexity by introducing datasets not used elsewhere (PATSTAT and Gateway to Research)

7. Relationship between exposure and business failure? What effect do the following exposure indicators have on the levels of business failures?

    (a) High industry / place exposure

(b) 'Is innovative?' flag - (if innovative companies were found to be more/less likely to fail)

(c) Presence of e-commerce / new strategy or technology adopted

(d) Covid notice containing “negative” information/outcome

(e) Social media update patterns

(f) Poor diversification opportunities

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

9. Indirect impacts - OUT OF SCOPE

   Placed out of scope due to data quality issues with the Glass network data, and a weak rationale for this task purely based on I/O tables data.

10. Detect business failures. Can we do it in a timely manner?

    (a) Explore the lag and accuracy of companies house [DATA ACCESS ISSUES]

    - **We now know this is likely to be >3 months**

    (b) Explore alternate datasets such as:

    - `https://www.gov.uk/government/statistics/incorporated-companies-in-the-uk-`
    - `https://www.gov.uk/government/statistics/monthly-insolvency-statistics-jun`

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.

11. How to deal with multi-site firms

    (a) How to resolve conflicting information between trading address (Companies House) and the addresses (up to 5) found by Glass AI.

    (b) 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)

    (c) What if there are more than 5 sites?

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

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.

1.3 Engineering "tasks"

14. Data modelling of Glass datasets, including the consistent merging of multiple snapshots

15. Fuzzy-matching

    (a) Develop reusable pipeline

    (b) Consider how to validate accuracy of fuzzy-matching

    (c) Choose precision-recall trade-off

16. Finding social media links - requires Nesta to scrape websites of organisations. NOTE: legal ramifications of scraping facebook data limit us to using Twitter only.

    (a) Do two small pilots:

        i. Scrape and retrieve twitter handles from business websites in location X - e.g. one local authority

        ii. Scrape and retrieve twitter handles from business websites in industry Y - e.g. one SIC code

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

    (a) Explore 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

(b) Fetch and store data from Twitter API

(c) Generate update statistics for each profile - i.e. ignore textual content

(d) Assess utility of data initially based on coverage and frequency

(e) Analyse textual content of twitter data [OUT OF SCOPE]

1.4 Task graph and proposed schedule

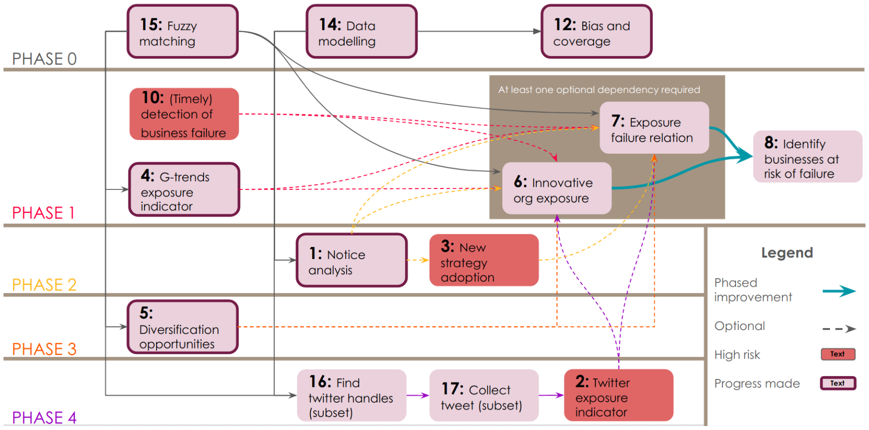
PHASE 0
15: Fuzzy matching
14: Data modelling
12: Bias and coverage
At least one optional dependency required
10: (Timely) detection of business failure
7: Exposure failure relation
8: Identify businesses at risk of failure
4: G-trends exposure indicator
6: Innovative org exposure
PHASE 1
1: Notice analysis
3: New strategy adoption
Legend
Phased improvement
PHASE 2
5: Diversification opportunities
Optional
High risk
Text
PHASE 3
Progress made
Text
16: Find twitter handles (subset)
17: Collect tweet (subset)
2: Twitter exposure indicator
PHASE 4

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

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.

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.

1.4.1 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.

1.4.2 Phase 1 - MVP

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

1.4.3 Phase 2 - Iteration

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

1.4.4 Phase 3 - Iteration

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

1.4.5 Phase 4 - Iteration

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