Using text data to improve industrial statistics in the UK

An exploratory study

Alex Bishop Juan Mateos-Garcia George Richardson

11 May 2021

Contents

Executive summary **2**

1 Introduction **3**

2 Data collection and processing **6**
2.1 Glass.ai 6
2.2 Labelling business websites with SIC codes by matching to Companies House 6
2.3 Pre-processing of Glass business descriptions 10

3 Using predictive modelling to analyse the SIC taxonomy **12**
3.1 Modelling methodology 12
3.2 Model performance and analysis 16
3.3 Conclusion 22

4 Hierarchical Topic modelling **23**
4.1 SIC similarities 24
4.2 SIC heterogeneity 25
4.3 Conclusion 30

5 Pilot strategy to develop a bottom-up industrial taxonomy **30**
5.1 Methodological narrative 31
5.2 Text preprocessing 32
5.3 Topic modeling 33
5.4 Sector reassignment 35
5.5 Postprocessing 38
5.6 Exploratory analyses 40
5.7 A hierarchical taxonomy 43
5.8 Conclusions and next steps 46

6 Conclusion **46**
6.1 Limitations of the current taxonomy 47
6.2 Advantages of a new, bottom-up taxonomy 47
6.3 Challenges for developing a new taxonomy 47
6.4 Next steps 48
6.5 Implementation considerations 48

Bibliography **49**

Executive summary

A growing number of economic policy agendas in the UK demand accurate, detailed and timely statistics about the industrial composition of the economy. This requires an industrial taxonomy that reflects existing industries and emerging ones enabling accurate classification of companies into sectors to produce economic indicators that can be used to inform policy.

The current version of the Standard Industrial Classification (SIC) that serves this purpose has some important limitations that could limit its usefulness. They include lack of timeliness (the taxonomy was last updated in 2007), presence of uninformative codes and difficulties accommodating companies that operate across multiple sectors.

In this report we use novel data sources and state-of-the-art machine learning methods to evidence the limitations of this taxonomy and explore options to develop a complementary taxonomy overcoming those limitations. In order to do this, we:

1. Match 1.8 million business website descriptions procured from Glass, a big data business intelligence company, with Companies House in order to label these descriptions with the SIC codes that these companies selected when they became incorporated
2. Use pre-trained text classification models to analyse the extent to which it is possible to predict companies' SIC codes based on their description, and potential explanations for model error that might be suggestive of limitations in the SIC taxonomy.
3. Use hierarchical topic modelling to characterise the homogeneity or heterogeneity of different SIC codes potentially helping us to identify codes that are particularly suitable candidates for decomposition into a more granular set of industries in a bottom-up, data-driven way
4. Pilot a strategy to build such a bottom-up industrial taxonomy by using the text of company descriptions to cluster them into 'text sectors.'

The results confirm our priors about the limitations of the SIC taxonomy currently in use: we identify important mismatches between company descriptions and the SIC4 codes where they are classified, and demonstrate the heterogeneity

of "other activities not elsewhere classified" codes that in some cases contain activities ranging from plumbing to social services and religious activities (in SIC4 8299) and from legal services to the organisation of clinical trials and the supply of renewable energies (in SIC4 7490).

Our emerging results also illustrate the potential of a bottom-up industrial taxonomy for generating sectoral categories that can be used to measure notable economic activities for example related to sustainability and the green economy. Some limitations of our approach include insufficient coverage of sectors outside the 'knowledge economy' and the presence of some noisy sectors in the data.

We conclude by setting out potential avenues to overcome these limitations and deploy, in a forthcoming ESCoE project, a bottom-up industrial taxonomy to analyse the UK economy in a way that demonstrates the value added of the methodology we have piloted here.

1 Introduction

The Standard Industrial Classification (SIC) provides an organising framework for the analysis of the sectoral composition of the UK economy (Hughes et al. 2009). It consists of a hierarchy of industrial codes that describe the economic activities of firms at increasing levels of resolution (see table 1 for an example).

Table 1: Example of the structure of the SIC taxonomy

Section	Division	Group	Class
Financial and Insurance Activities (K)	Financial service activities, except insurance and pension funding (64)	Insurance (651)	Life insurance (6511)

Companies self-select their SIC code when they register with Companies House. Subsequently, they may be reassigned into a new code if, for example, a mistake is identified while the company undertakes an official business survey.

Eventually, these classifications are used to produce official statistics about the business population and the sectoral distribution of productivity, employment, occupations and salaries through a variety of ONS products such as the Interdepartmental Business Register (IDBR), the Annual Business Survey (ABS), the Business Register Employment Survey (BRES), the Annual Population Survey (APS) or the Annual Survey of Hours and Earnings (ASHE).

The widespread use of the SIC taxonomy to produce 'sectoral cuts' of other economic statistics underscores its importance: different sectors vary in their

productivity, geography, skills needs, innovation activities, business models and internationalisation among other factors so understanding their levels of activity and how it evolves over time is critical for informing a host of economic policies. The increasing importance of industrial policy, regional rebalancing ('levelling up') and sustainability policy agendas places an additional premium on access to timely and granular statistics about the industrial composition of the UK and its national, regional and local economies.

There is increasing awareness of the limitations in the version of the SIC taxonomy currently in use which makes it less relevant for these policy needs (Hicks 2011). They include:

- Lack of timeliness: The version of the SIC taxonomy currently in use was last updated in 2007, making it unsuitable for the analysis of industries that have emerged since then and are of particular interest for policymakers such as Artificial Intelligence, Fintech, Renewables and the "Gig economy."
- Presence of uninformative sectors: Out of the ca. 600 four-digit SIC codes ('classes') in the SIC taxonomy, 52 refer to 'other activities' or 'activities not elsewhere classified,' employing 15% of the workforce in 2019 according to BRES.
- Difficulties accommodating companies that straddle sectors: the SIC taxonomy is completely exhaustive and mutually exclusive, meaning that all companies are classified in a code and each company is classified in a single code. This has the advantage of avoiding double counting but might create challenges classifying business that undertake activities captured in several SIC codes, such as for example a fintech company that applies digital technologies (captured in Division 62:'Computer programming, consultancy and related activities') in financial services (captured in Division 64:'Financial service activities, except insurance and pension funding').[1]
- Misclassification: Together, all the reasons above lead to concerns that companies might select the wrong code because they do not see their activities reflected in the current taxonomy or opt for a "not elsewhere classfied code" even when a suitable code is available somewhere in the SIC taxonomy.[2]

The increasing availability of open and web data about what companies do (or say they do) provide some interesting opportunities to address some of these limitations in the current SIC taxonomy (Bean 2016).

In particular, web sources such as business websites and sector-specific directories and web portals have been used to measure the digital economy (Nathan and Rosso 2015), the video-games sector (Mateos-Garcia, Bakhshi, and Lenel 2014), the 'immersive economy' (including technologies such as Virtual Reality and

[1] Here, it is worth noting that businesses can select multiple SIC codes when becoming incorporated but it is unclear how often they do this, and information about business secondary codes are rarely reported.

[2] As we noted previously some of these instances of misclassification may be rectified subsequently when additional data is collected, for example through a business survey.

Augmented Reality) (Mateos-Garcia, Stathoulopoulos, and Thomas 2018) and industrial clusters in the UK (Nathan et al. 2017). In a previous analysis, we used business websites to map the geography of businesses using emerging technologies in the UK (Bishop and Mateos-Garcia 2019).

This approach has the advantage of relying on what businesses say that they do in order to reach their customers instead of their engagement with the administrative process of selecting a code when they become incorporated. We would expect these statements to be more timely than the codes in the SIC-2007 taxonomy insofar businesses have incentives to update their websites regularly, and to mention new products, services, processes and technologies that may be of interest to their customers, thus capturing emerging industries. They should also reflect the multiple economic activities that businesses engage in regardless of whether they are confined to a single industrial code.

Unstructured descriptions of business activities are however not without their limitations. Novel sources may capture unrepresentative samples of the business population, and business descriptions are likely to be noisy, in part because they serve the purpose of promoting products and services and attract new customers rather than providing an accurate description of what businesses do. There is also the significant challenge of transforming all this unstructured information into a taxonomy that can be used to measure economic activities in different industries and eventually make policy-relevant statements about the composition of the economy.

In this working paper we report emerging findings from an exploration of the opportunities and challenges for building a bottom-up industrial taxonomy based on text data complementing the SIC taxonomy and addressing some of its limitations.

In order to do this, we match a dataset of business website descriptions obtained from Glass, a business intelligence startup, with Companies House. We then assess the alignment between SIC codes at the four-digit level and business descriptions using supervised and unsupervised machine learning methods, and pilot an approach to build a bottom-up industrial taxonomy based on an analysis of the text in company descriptions.

The structure for the report is thus:

Section 2 introduces our data and how we have processed it with a special focus on the fuzzy matching algorithm we have developed in order to combine business website descriptions with Companies House and the natural language processing pipeline we use to process company descriptions.

Section 3 presents the results of a supervised machine learning analysis where we train a predictive model on our labelled dataset in order to determine the extent to which it is possible to predict 4-digit SIC codes using the text in business descriptions, and the explanation for various instances of misclassification.

Section 4 presents the results of an unsupervised machine learning analysis where

we train a hierarchical topic model on our dataset with the goal of assessing the semantic homogeneity / heterogeneity of 4-digit SIC codes (i.e. the extent to which they contain companies with widely varying descriptions of their activities) as well as semantic overlaps between codes in different parts of the SIC taxonomy.

Having 'dissected' the SIC taxonomy using machine learning methods and diagnosed some of limitations, Section 5 trials an experimental, iterative strategy to build a bottom-up industrial taxonomy that addresses some of them. In order to do this, we use a network-based topic model to decompose 4-digit SIC codes into more granular text sectors. We explore the opportunities that this opens up for decomposing uninformative SIC codes into their constituent parts, studying policy-relevant 'sectors' such as the green economy, characterising more accurately the composition of local economies, and clustering text sectors into a hierarchical industrial taxonomy.

Section 6 presents conclusions and next steps.

2 Data collection and processing

2.1 Glass.ai

The core dataset for our analysis has been obtained from Glass, a startup that uses machine learning to collect and analyse business website data at scale. More specifically, Glass begin from the universe of UK websites in web domain registers, identifies those that are highly likely to belong to a business, and extracts relevant information about them including their description, postcodes and sector based on an industrial taxonomy developed by LinkedIn. In this project, we work with information about 1.8 million business websites (which according to Glass account for 90% of UK business websites) collected in May and June 2020.

The granular business descriptions contained within this dataset can be used to understand a business' economic activities at a higher level of resolution than is possible using the SIC taxonomy.

2.2 Labelling business websites with SIC codes by matching to Companies House

In order to obtain SIC codes for business websites we match businesses in the Glass dataset to the Companies house business registry. We use the monthly data snapshots for May, June, and July 2020 - available at the time from the Companies house website - as this corresponds to the period of time for which we have Glass data.

The matching methodology matches the names of companies in Companies House with the names extracted by Glass from business websites based on their similarity according to some measure[3]. Naively comparing the similarity of all combinations (~4 million Companies House companies x ~1.5 million Glass websites) of names is computationally infeasible - this would take roughly 20 years (on a single CPU) assuming we could do 10,000 similarity computations per second. This leaves us with three options, with the third being the only one that is reasonable and possible:

1. Performing the computations on a supercomputer
2. Make a breakthrough in the field of computing by improving the performance of a fundamental algorithm by several orders of magnitudes
3. Reduce the number of pairwise comparisons by only comparing pairs that are 'likely' to be matches

It may sound paradoxical that we could identify pairs that are likely to be matches without calculating their similarites up-front. We achieve this by using two approaches: probabilistic data structures - namely the Minhash combined with Locality Sensitive Hashing (LSH) - and the cosine similarity computed using a chunked dot-product.[4]

2.2.1 Probabilistic data structures (PDS) approach

Convert company names to a set of 'k-shingles' - a sliding window k characters. For example, for $k = 4$ "acme co" would become the set {"acme," "cme," "me c," "e co"}.

One can measure the similarity between the k-shingle sets of two company names with the Jaccard similarity -

$$J(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|}$$

.

The Jaccard similarity can be approximated using Minhash and LSH - a method originally used by search engines to detect near-duplicate web pages to improve their search results (Broder et al. 2000). Below we outline the high-level concepts behind the approach. The interested reader should refer to Chapter 3 of (Rajaraman and Ullman 2011) for a thorough treatment of the theory and implementation behind MinHash and LSH.

A Minhash is cheap to compute and has the following relation to the Jaccard similarity: The Jaccard similarity of two sets is the probability that their minhash is the same. By computing many (n) minhashes of each name (second column

[3] For example the Levenshtein distance (Levenshtein 1966).

[4] This approach does in fact compute similarities up-front; however the similarity measure used is cheap to compute.

in table 2) we can approximate probability that the minhash of two sets is the same and thus their Jaccard similarity - as $n \to \infty$ this will become exact. Next we use Locality Sensitive Hashing and group the n MinHashes of each name into b bands of size r ($n = br$) and checking for collisions in each band (third column table 2) we identify names with a high probability of having the same MinHash. Putting this all together, for each name in the first dataset of names (e.g. for each Glass name) we can efficiently identify names in the second dataset of names (e.g. Companies House names) that have a high probability of having the same MinHash and thus are likely to be highly Jaccard similar.

Table 2: Example of MinHash and LSH on three company names (first column). 6 different Minhashes are computed for each name (second column) and then grouped into two different bands and checked for collisions (third column). There is a hash collision between the first two names in the first bucket, therefore they are identified as similar.

Name	$n = 6$ Minhashes of Name	LSH Groups ($b = 2, r = 3$)
"acme co"	`[1,1,2,4,3,2]`	`[112, 432]`
"acme company"	`[1,1,2,3,3,2]`	`[112, 332]`
"accenture"	`[2,2,2,3,3,1]`	`[222, 331]`

2.2.2 Chunked cosine similarity approach

We augment the pairs of likely similar names identified by the PDS approach with pairs identified by: 1. count-vectorising names to produce a matrix for each set of names (rows are names, columns are tokens); 2. {Term frequency}-{inverse document frequency} (TFIDF) transforming our matrices (considering both datasets of names to come from the same corpus for this purpose); 3. computing the cosine similarity of the TFIDF scores of company names and identifying the top n most cosine similar pairs in the second dataset for each name in the first dataset.

This is performed in a computationally efficient manner by performing the dot-product between TFIDF matrices in a chunked manner.

This provides a complementary approach to the PDS approach, which excels at capturing small character level discrepancies, by taking into account word ordering differences and information about the frequency of words across the corpus of company names.

2.2.3 Computing exact similarities of subset

After identifying these sets of similar names we can apply the exact similarity measures that were computationally infeasible to naively apply across the whole

dataset.

The similarity measures we chose to use for matching Glass to Companies House were:

- Jaccard Similarity of 3-shingles (exact)
- Cosine Similarity of TFIDF scores
- Levenshtein distance

2.2.4 Choosing 'best' matches

After computing exact similarities for our subset of pairs, we choose the 'best' matches by identifying the Companies House name with the highest mean similarity score for each Glass name. Each Glass organisation only appears once but each Companies House organisation may appear multiple times. We only consider matches with a similarity score of 75% or higher - which we empirically determined as a sensible threshold.

2.2.5 Criticisms

This matching is not exact and problems do exist. We have split the discussion of this matter into problems with the algorithm chosen to perform the matching and problems with the datasets themselves.

Problems with the algorithm There are a number of hyper-parameters such as the shingle-size (k) and the number of MinHashes (n) to choose which may affect the accuracy of this approach. In particular, choosing n too low may introduce false negatives (we miss matching pairs); however larger values n are prohibitive due to memory consumption.

Furthermore, choosing the best match based on the mean similarity is a fairly naive approach which one could replace with better heuristics that take into account factors such as string length, or other data such as the geographic proximity of two businesses (noting that there can be multiple Glass addresses for a website and that these are not guaranteed to match-up to the registered addresses present in Companies House).

Problems with the datasets

Glass The business names in the Glass data are extracted from the text of websites and therefore are not guaranteed to be extracted correctly or even to correspond to the officially registered name of a company within Companies House.

Companies House Due to the nature of Companies House, some matches may be to the wrong part of a conglomerate company which may have a different SIC designation. Furthermore, many companies in Companies House have inaccurate SIC designations. The IDBR team within the ONS have a modified version of the Companies House dataset which reassigns SIC codes and may contain information about company groupings. Due to the timescales of this project it was not possible to access this data.

2.3 Pre-processing of Glass business descriptions

The analysis of section 4 and section 5 requires processing the raw descriptions extracted from business' websites by Glass into a form we can use to, for example, train a topic model on company descriptions or generate their vector representation to measure similarities between companies. We are specially interested in removing text which is uninformative about the industry where a company operates in but is likely to appear in a website, such as for example its location.

In order to do this, we build a Natural Language Processing (NLP) pipeline using the Spacy and Gensim Python libraries, which convert the raw 'string' of a business website description into an ordered list of 'tokens' where each token is a unigram, bigram or trigram composed of the lemmatised form of a word or an (uppercased) entity type label (for a subset of entity types) (**vrehuuvrek2011gensim?**).

For example: `"I went to the Eiffel tower and visited my friend Jacques"` -> `["went", "GPE", "visit", "friend", "PERSON"]`

Steps 1-8 are performed or rely on information extracted using the Spacy `en_core_web_lg` model, whilst steps 9-10 are performed or rely on information extracted using Gensim.

1. Tokenisation
2. Named entity recognition (NER) - Predict named entities using a transition-based method
3. Lemmatise - Assign base forms to tokens
4. Merge entities - Merges series of tokens predicted by Spacy to be an entity into a single token
5. Filter
    - stopwords
    - punctuation
    - whitespace
6. Extract to list of strings - lemmatise or entity form

7. Generate n-grams - Use `gensim` to generate bi-grams, requiring that a bigram occurs at least 10 times and that the normalised pointwise mutual information (NPMI)[5] is 0.25 or higher.

8. Filter

- short tokens
- combinations of stop words
- Words with low and very high frequency (those occurring less than 10 times and in more than 90% of documents)

Token lemmatisation/remapping In step 8 tokens that are entities in the following categories are renamed to correspond to their entity category (upper-cased):

- CARDINAL
- DATE
- GPE
- LOC
- MONEY
- NORP
- ORDINAL
- ORG
- PERCENT
- PERSON
- QUANTITY
- TIME

We hypothesised that replacing these entities with their entity type name helps keep more information in the bag of words representation, particularly when entity types can be formed into n-grams with other words.

The alternative is that individual dates, people, organisation names etc. are too infrequent in the corpus to contribute information to the topic modelling approach.

Several entity categories such as `WORK_OF_ART`, `LANGUAGE`, `LAW`, `EVENT` were excluded as an empirical assessment of the classifications on sample business descriptions appeared inaccurate.

Furthermore, `PRODUCT` was left out because in this problem context (generating an industrial taxonomy) this is valuable information that we do not wish to homogenise.

Tokens that are not entities are replaced with their lowercase lemmatised form.

[5] NPMI $\in [-1, 1]$ where a value of: -1 indicates tokens never occur together; 0 indicates independence; and 1 indicates complete co-occurrence.

3 Using predictive modelling to analyse the SIC taxonomy

How well can a machine learn the relationship between what a company says it does and the SIC code that company has been assigned?

In this section, we attempt to answer this question by training a machine learning model to predict 4-digit SIC codes based on company descriptions and analysing the results. We do this in order to better understand the challenge of fitting companies into the SIC-2007 taxonomy based on the activities that they use to describe themselves, and identify opportunities to improve classification through the bottom-up taxonomy we develop later in the paper.

To do this, we have taken advantage of recent advances in machine learning, in particular neural network architectures known as transformers, which are particularly well-suited to natural language tasks. Transformers are capable of modelling linguistic relationships in ways that make them suitable for research and applications in text classification, question and answering, text summarisation, machine translation and text generation. They have been rapidly adopted by the open source community, with tools such as the Transformers Python library allowing developers and researchers to easily obtain and use models for various natural language processing (NLP) tasks (Wolf et al. 2020).

Importantly for this work, the library provides access to *pre-trained* models, which have already been trained on a large corpus of text to recognise language patterns and can subsequently be adapted for new tasks with relatively little data. Here we describe first how we have used such a model to predict SIC codes and second, how we have used machine learning performance metrics to understand where the model succeeds and struggles, which is informative about the limitations of the SIC taxonomy and how it is used by companies when they register in Companies House.

3.1 Modelling methodology

3.1.1 Transformers

We build a multi-class classifier that attempts to predict one of the 615 4-digit SIC codes, using the description of a company as its input. The development of the classifier is treated as a supervised machine learning problem - one in which a labelled dataset is used to train a model. As we have described above, there are no official datasets that contain company descriptions and their associated SIC codes. In this case, we have used the fuzzy matched data from Glass AI and Companies House to obtain a labelled dataset where SIC codes are associated with business descriptions.

We use a transformer model to construct the classifier. Transformers and other neural networks are in large part able to learn complex linguistic relationships because of the large numbers of parameters that they contain. These parameters are the mathematical weights and biases that describe the relationships between different nodes inside the network and that ultimately determine the relationship between the inputs of a model and its outputs. Transformers can have many millions or billions of parameters to tune, which in turn requires large volumes of data and ultimately means that even with optimised software and hardware, training a model is time and energy intensive.

Fortunately, researchers have discovered that once a model has been trained on a large volume of data for a particular task, the patterns that it learns are often highly generalisable. A transformer that is trained on a large corpus, such as the text from English language Wikipedia, for one task will result in a model with parameter values that represent a large number and diversity of patterns that are valuable for other NLP tasks. Because of this, training a model from scratch is often unnecessary and instead, developers can opt to "fine-tune" an existing model for their specific problem.

The Transformers library makes this particularly easy by offering a consistent tool for working with models built in different frameworks (TensorFlow and PyTorch) and with different model architectures, enabling access to models that have been pre-trained on large datasets with accelerated hardware. This allows community members to share models with each other, reducing the amount of collective time and resources that must be spent on model training. It also provides an interface for fine-tuning those models.

For the SIC code classifier, we fine-tuned an English language DistilBERT model (Sanh et al. 2019). This is a transformer that has a reduced number of parameters to reduce disk space, memory use and training times, but that still performs comparably to other larger models. The model was trained on a masked language modeling task where 15% of words in a sentence are randomly masked and the model must predict what those masked words are based on the unmasked content. This gives the model an inner representation of the English language that is then transferrable to a classification process. The additional component to turn the model into a classifier is the addition of a fully connected, final layer to the neural network that has a number of nodes equal to the number of classes being predicted (in our case 615 SIC codes).

The entire model is then trained by feeding it the text examples in a dataset, requiring it to reduce the loss (error) between the predictions made in the final layer and the labels in the dataset. This process tweaks the values of all the parameters in the network, as well as the final classification layer.

3.1.2 Implementation

In practice, several design choices were made in the development of this model. The first is that a higher match threshold between the Glass AI and Companies House datasets was chosen compared to previous work we have done with this data. In an initial prototyping of the model to predict the highest level of the SIC index (the Section), we noticed that predictions on company descriptions with a higher matching score yielded a higher F1 score.[6] The micro F1 score peaked when the match threshold was around 75, therefore only company descriptions that matched to Companies House with a minimum score of 75 were chosen to train, validate and test the model. In total, after the matching threshold had been applied, there were around 350,000 samples for training. In this phase of development, 80% of these were used for training, while 20% were held out as the test set for analysis.

Second, an optimisation technique was applied to mitigate a limitation of transformers that can slow down the training time for the fine-tuning process. Transformers are trained by passing data through the model in batches. Several examples of text are shown to the model before its predictions are evaluated against the true labels and the parameters are updated. In each batch, the number of input features - in this case, the input features are the tokens in the company descriptions - must be consistent across all samples. Because of this, shorter texts in a batch are padded up to the maximum length with a filler token that has no bearing on the model parameters.

In addition, transformers also have a maximum input size (for DistilBERT this is 512 tokens). Any examples in a batch can only be padded up to this length or a shorter length specified by the user. Any sequences longer than the specified maximum or upper limit are truncated. This is relevant to the training speed as the time it takes for a transformer to process a piece of data is quadratic with its length. For any training process with variable length texts there is always the challenge of setting a maximum length that captures sufficient volumes of information in longer sequences and that does not impose a significant speed penalty. Fortunately, the batching process permits the use of dynamic padding and uniform batching, two techniques that when combined can offer a significant speed up.

Dynamic padding is the process of applying tokenisation at the point of batching the data, rather than on all data at the start of the model development process. In this way, a maximum length can be chosen that is no longer than the longest sequence in a batch. This helps by ensuring that no examples are padded more than they need to be. A global maximum can still be set to ensure no sample is padded beyond a certain length.

Uniform batching is the approach of grouping texts of similar length together

[6] The F1 score is the harmonic mean of a model's precision and recall i.e. its ability to predict true positives while avoiding false negatives.

before they are passed to the transformer in batches. In this way, the amount of padding needed in a given batch is reduced yet again. One way of doing this is by sorting the data by length to ensure that shorter texts are found close to each other. A schematic for dynamic padding and uniform batching is shown in figure 1. Both methods were applied to the data during the training process for the SIC classifier and resulted in significant speedups.

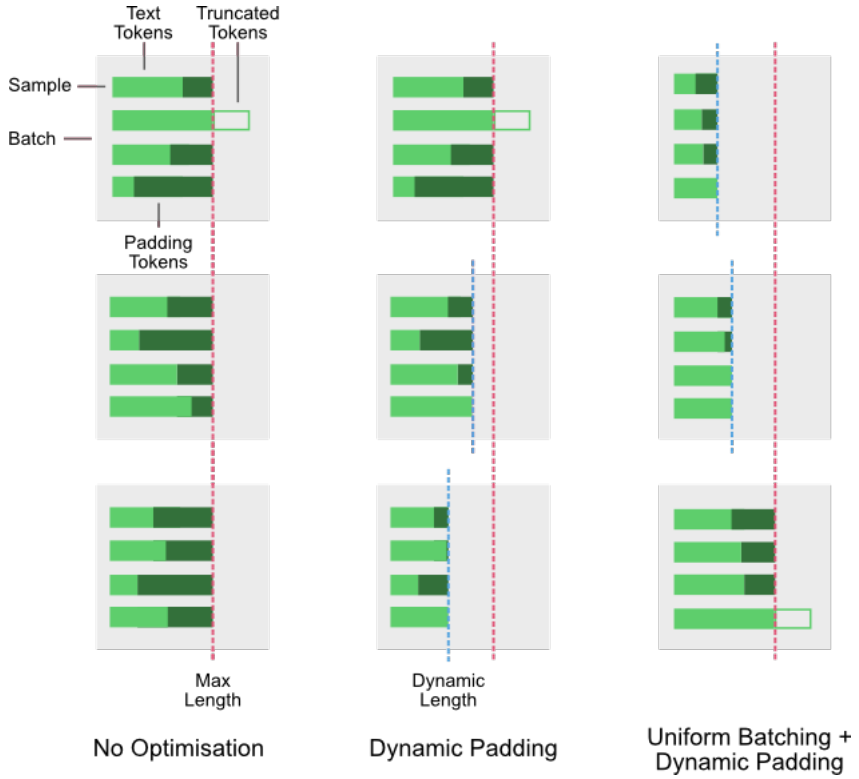

Figure 1: With dynamic padding and uniform batching, texts of similar length are grouped together and padded only to the length of the longest sequence, ensuring that all batches, and therefore all sequences, take a minimal time to process.

After training the model by following the steps above, the classifier was applied to the remaining 20% of company descriptions in the test dataset. These predictions form the basis of the analysis in the following section.

It is worth noting that the process described here presents some limitations, particularly in relation to the training data.

1. The dataset has a high class imbalance. There are some SIC codes with many thousands of examples and others with only a handful. This means that the information available to the model for learning some SIC codes is

significantly larger than for others.

2. We cannot consider the labelled dataset as a gold standard, as manual inspection of the data shows that in some cases the label associated with a company description does not constitute a reasonable fit. In yet other cases, the description provided is not adequate to easily determine a single suitable label out of the 615 available, as a combination of codes would more suitably describe some organisation's activities. However, in some ways it is these aspects that make this analysis interesting.

3.2 Model performance and analysis

Evaluating a model with 615 possible output classes is not a trivial task. Machine learning often involves making performance trade-offs both within and between classes and therefore the number of possible classes increases the complexity of this task by making the final optimisation objective more difficult to define. In this analysis we do not attempt to comprehensively determine the efficacy of the model, but rather make use of various metrics typically used within model evaluation to highlight findings that are relevant to understanding the limitations of the SIC-2007 taxonomy and informing the creation of a data driven alternative.

The first observation is that the classifier only makes predictions for 95 of the possible 4-digit SIC codes, despite the training data covering companies from all of the 615 codes. Classification metrics, such as precision, recall and F1 score are therefore only available for a subset of the codes, however we do shed some light on why this extreme aspect of low performance occurs by combining these with other metrics.

In figure 2 we see the most performant SIC codes according to their F1 score. We can see that there are only 6 codes with scores above 0.7, before a dropoff and gradual decline in scores. From a manual inspection of the results, there appear to be no obvious characteristics of the codes, such as industry type, that determines their performance, however it is worth noting that there are very few low-specificity codes (such as those that are defined by their inclusion of n.e.c. - not elsewhere classified) among the most performant. As we will see, the results are more nuanced than this. Overall, this highlights that the model struggles to assign companies a single label succesfully according to the test data.

While this gives an overall impression of accuracy, it does not tell us much about the nature of the misclassifications. The set of companies with a particular SIC code may be misclassified into one other class or many, and these may relate closely to the original category or be very distant in terms of industrial activity. In order to determine the diversity of misclassifications for companies in each SIC code and to understand why this might be happening, we calculate two additional metrics: the Shannon index and the Silhouette score.

The Shannon index is a measure of entropy, often used to describe the diversity

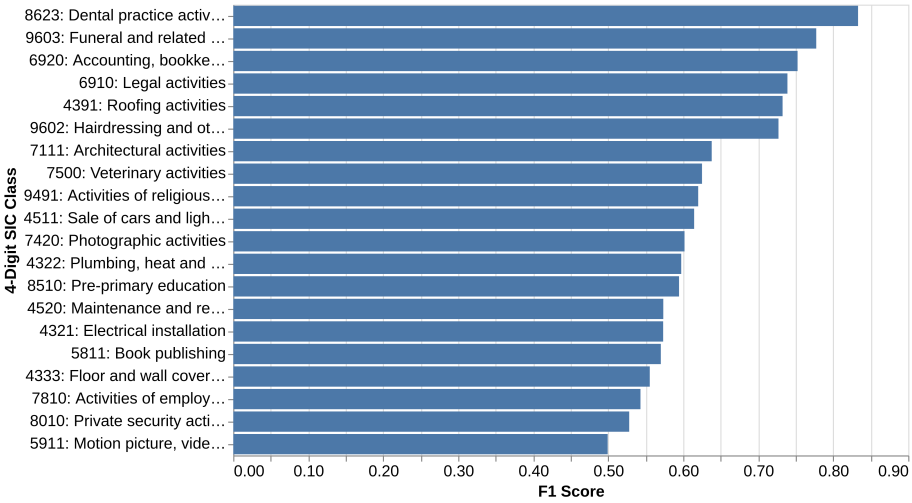

Figure 2: Top 4-digit SIC codes according to their F1 scores.

of states in a system. For companies in the test set with a SIC code, i, it is defined as

$$H_i = -\sum_{j=1}^{s} (p_j \log_2 p_j) \tag{1}$$

where s is the number of operational taxonomic units (in this case 615) and p_j is the proportion of companies classified with SIC code j. This tells us the degree to which companies are misclassified into the possible codes.

Second, we calculate the silhouette coefficient for each SIC code. The coefficient is a metric that describes how well clustered a system is. That is how close points in a cluster are to other points in that cluster, as opposed to being close to points belonging to other clusters. In our case, we do this based on the company descriptions to understand whether companies labelled with a SIC code are clustered together with other companies with semantically similar descriptions, or whether they are dispersed in poorly defined clusters.

To generate this value, we must first project the companies into some numerical space that represents the semantics of their descriptions. We use another pre-trained DistilBERT based transformer model for this task. We choose a version of this model that has been trained on semantic similarity tasks and without any further fine-tuning we process the company descriptions to produce dense 768 dimensional vector representations of the texts.

With each of the companies now occupying some point in semantic space, we calculate the mean sample silhouette coefficient for each SIC code. The sample

silhouette score for a single company, $s(x)$, is defined as

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

where $a(x)$ is the distance of a company description from all other companies with the same 4-digit SIC label and $b(x)$ is its distance from all other companies. We use the cosine distance as our distance metric. We then take the mean of the sample silhouette coefficient for all companies that are labelled with a 4-digit SIC code in the test set.

By comparing these two metrics we can see the degree to which companies within a code having dispersed descriptions determines the range of predicted SIC codes. Indeed, figure 3 shows that a SIC code with a higher silhouette coefficient tends to lead to less diverse predictions. This makes intuitive sense as SIC codes whose company descriptions are less tightly clustered will necessarily have some overlap with other codes. This will result in a challenging situation for the classifier when it comes to learning and predicting for company descriptions that are in dispersed clusters or are on the periphery of their group. Almost all of the SIC codes have a silhouette score below zero, highlighting the degree to which this is the case. This means that in the majority of codes, company descriptions can be interpreted as more similar to companies in other industries than their own. It is also notable that there are some SIC codes which have a range of Silhouette scores but a Shannon index of zero, indicating that all companies in the test set with this code have been predicted into the same class.

However, a low Shannon index, driven by a high silhouette score does not necessarily mean that companies in that sector are well classified. In fact it may mean that companies in a SIC code with those characteristics are almost entirely misclassified into a single other code. A high Shannon index is also not necessarily a bad outcome, despite the fact that it points to apparent misclassification. For example, companies in a sector with a lower silhouette score might be resolved into a larger number of more appropriate sectors through the classifier's predictions.

In table 3 we can see that those sectors with a high Shannon index are those that are very non-specific and often uninformative, including several "n.e.c." codes. This suggests that many of the companies that were originally labelled with these sectors might be reclassified by the model into what it believes are more appropriate sectors.

Table 3: The top and bottom 10 4-digit SIC codes sorted by the Shannon index of their predicted codes. Only codes with non-zero Shannon indices are shown.

SIC4	Description	Shannon	Silhouette
9609	Other personal service activities n.e.c.	5.32	-0.29
8299	Other business support service activities n.e.c.	5.27	-0.28

SIC4	Description	Shannon	Silhouette
7010	Activities of head offices	5.08	-0.29
6420	Activities of holding companies	4.85	-0.27
4799	Other retail sale not in stores, stalls or markets	4.44	-0.3
8110	Combined facilities support activities	4.33	-0.24
7490	Other professional, scientific and technical activities n.e.c.	4.26	-0.26
4719	Other retail sale in non-specialised stores	4.24	-0.3
6820	Renting and operating of own or leased real estate	4.19	-0.26
6399	Other information service activities n.e.c.	4.18	-0.23
...	...	...	...
3091	Manufacture of motorcycles	0.81	-0.19
3220	Manufacture of musical instruments	0.81	-0.26
3512	Transmission of electricity	0.81	-0.24
2594	Manufacture of fasteners and screw machine products	0.73	-0.1
1071	Manufacture of bread; manufacture of fresh pastry goods and cakes	0.67	-0.15
2311	Manufacture of flat glass	0.65	-0.21
2219	Manufacture of other rubber products	0.62	-0.15
2363	Manufacture of ready-mixed concrete	0.59	-0.17
8623	Dental practice activities	0.59	0.08
1083	Processing of tea and coffee	0.39	-0.21

As the F1 score is the harmonic mean of precision and recall, it provides a view of model performance that attempts to balance these two measures of accuracy. This obscures the fact that for this task, there are codes which do not necessarily share the same level of performance across both metrics. To visualise the distribution of SIC codes across the dimensions of precision and recall, we plot the model performance according to these values in figure 4. The resulting distribution suggests that there are two notable modes of classification - those where either precision or recall dominates. Codes that have a high recall and a low precision will be the result of a high proportion of false positives, suggesting that companies in other sectors might have descriptions that semantically similar to a large number of companies within the code. Codes that have a high precision and low recall suggest that there are sectors which have a well-defined core of companies with semantically similar descriptions, but other companies which the classifier believes fall better into other categories.

In addition to the precision and recall positions of the codes, figure 4 also uses the size of the points to highlight the degree of semantic overlap between companies labelled with that code. We placed the DistilBERT embeddings described above into a FAISS index to perform an efficient nearest neighbours search in semantic space (Johnson, Douze, and Jegou 2019).

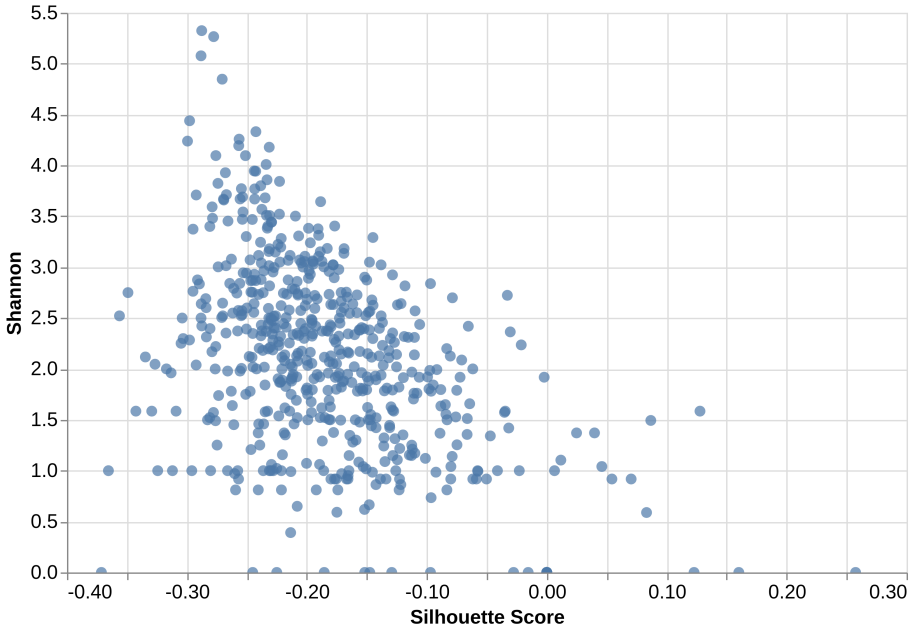

Figure 3: Comparing the silhouette score of company description embeddings and Shannon index of predictions for each 4-digit SIC code shows that SIC codes with less well defined semantic space are more likely generate a diverse array of predictions.

For each 4-digit SIC code, we calculated the proportion of companies whose nearest neighbour is also in the same code, which is shown on the chart. In general, we can see that a higher accuracy, in terms of either precision or recall, occurs in sectors where the intra-code semantic overlap is higher, however there are some exceptions. An extension here would be to calculate metrics that account for both the overlap and dispersity to analyse whether it is indeed clusters with a tight semantically congruous core that result in high precision and vice-versa.

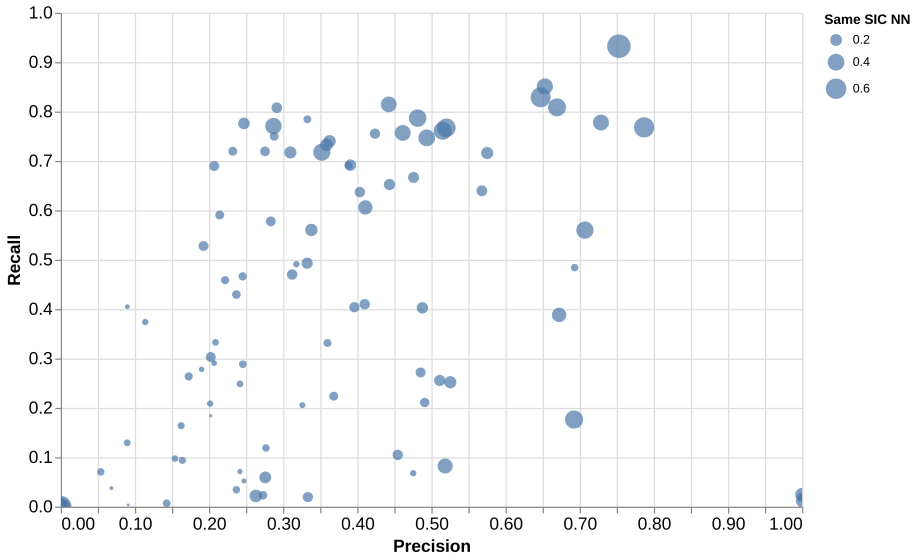

Figure 4: The precision and recall metrics show that some SIC codes are dominated by different classification characteristics, likely to be driven by the degree and type of semantic overlap with companies in other codes.

Overall, it is clear that there are sources of misalignment between many company's descriptions and the SIC codes that they are labelled with, and that lead to the prediction errors. A final demonstration of the impact of this is a visualisation of 4 of the less specific SIC codes across the semantic space of all company descriptions, namely 8299, 9609, 7022 and 7490. The 768 dimensional vectors are projected down to 2 dimensions for visualisation purposes via dimensionality reduction using the UMAP algorithm (McInnes, Healy, and Melville 2018) - see figure 5. Although it is not advised to draw conclusions about the exact relationship between two points according to their relative positions, due to fluctuations in the density of the space, some global trends can be determined. In this case, it is immediately clear that these codes are highly dispersed among companies with descriptions that cover the semantic space, and are neighbours to companies from a wide range of other sectors. This is likely to have a very large impact on the ability of the model to learn distinct patterns in the company

descriptions belonging to these sectors, but also the sectors they they overlap with.

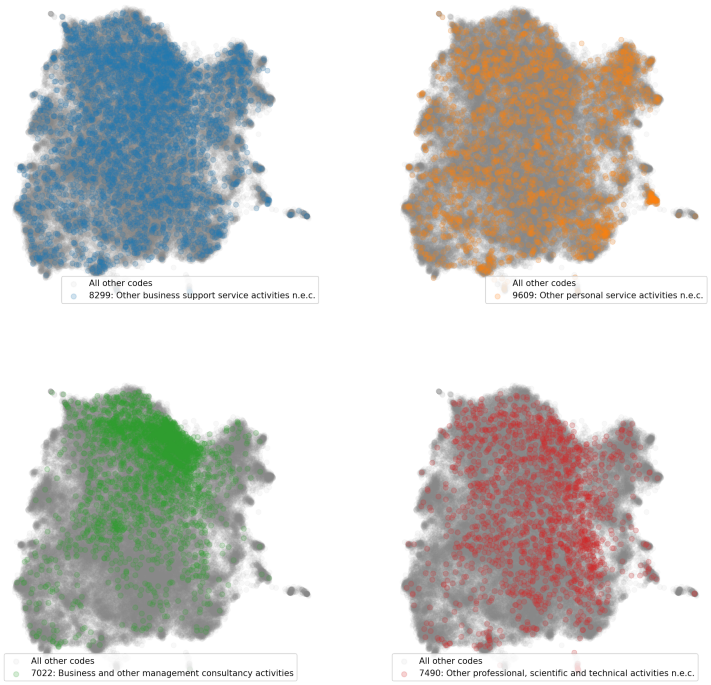

Figure 5: UMAP visualisation of the position of companies in sectors 8299, 9609, 7022 ans 7490 in vector space.

3.3 Conclusion

In this section we have demonstrated the challenge for state-of-the-art natural language classification to identify the appropriate SIC code for a company based on a labelled dataset. We show that this is in large part due to the degree and nature of semantic overlap of company descriptions within and between labels according to Companies House.

In some cases, we believe that this is a result of uninformative codes being

applied to companies that better belong to another, more specific sectoral label. In other cases, it is because a single label is inadequate to describe the company's activities as described by themselves on their business website. While it might be sufficient to use two or more existing codes to describe such companies, it might also be the case that a suitable code does not yet exist.

In other cases, the error may stem from companies being unintentionally mislabelled on Companies House due to a lack of guidance and clarity in the process of selecting an industrial code when establishing or updating a company's records. As one example, we found two companies in the dataset offering domestic and commercial electrical installations and testing with one labelled as belonging to 3512: Transmission of electricity and the other as 4321: Electrical installation.

In conclusion, the data does not permit our model to satisfactorily learn how to classify companies according to their descriptions. This is a result of both typical limitations encountered in machine learning, such as class imbalance, as well as the nature of the SIC-2007 taxonomy and its application to companies in the UK. This motivates and informs our pilot to develop a bottom-up industrial taxonomy in Section 5.

4 Hierarchical Topic modelling

In this section, we present an unsupervised machine learning analysis where we train a hierarchical topic model on the processed Glass descriptions. This allows us to assess the semantic heterogeneity of 4-digit SIC codes (i.e. the extent to which they contain companies with widely varying descriptions of their activities) as well as semantic overlaps between codes in different parts of the SIC taxonomy.

The hierarchical topic model approach used is the TopSBM (Gerlach, Peixoto, and Altmann 2018) model. This approach confers multiple advantages over the more traditional Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) frequently used in the literature such as automatically selecting the number of topics; yielding a hierarchy of topics; permitting a more heterogeneous topic mixture than is permitted by LDA; and, crucially for the analysis of section 4.1 generating document clusters. Theese benefits are not without cost - due to the high memory use[7] of this methodology we use only 100000 Glass descriptions (those with the highest match score to Companies House are selected) to fit the model.

[7]Fitting the model on 100000 documents required use of a machine with 64GB RAM.

Table 4: Number of topics and number of document clusters for top 4 levels of the fitted model's hierarchy.

Level	Number of topics	Number of clusters
0	384	370
1	74	56
2	11	11
3	2	3

4.1 SIC similarities

Figures 6, 7, 8 show the cosine similarity between SIC divisions calculated using the three most granular levels (table 4) respectively of our hierarchical topic model.

The similarities are calculated by aggregating documents into SIC codes (division level) and calculating the cosine similarity between the document cluster distribution of divisions. We choose the division level to discuss this component of analysis at because it is the most granular level of the SIC taxonomy at which it is feasible to visualise and compare pairwise similarities. The most appropriate level of our model hierarchy to analyse SIC similarity at is more subjective - lower levels pick out very strong relationships which hold in the presence of a finer topic/cluster structure, whereas higher levels better pick out higher-order structure.

The diagonal block-structure (particularly visible in figure 8) corresponds to the SIC taxonomy structure; however there is also significant off-diagonal structure which highlights the richness of novel data-sources such as business website descriptions.

For example, divisions 10-15 (broadly the manufacture of food, beverage, and clothes) are highly similar to divisons 56 (food and beverage service activities) and divisions 46-47 (wholesale and retail trade). Many similar intuitive relationships exist across related extraction, manufacturing, and services industries in disparate parts of the SIC taxonomy.

Table 5 lists a few more intuitively similar SIC divisions which are not captured by the SIC taxonomy but are well-captured by the Glass data and topic modelling approach.

Table 5: High similarity pairs of divisions not captured by the SIC taxonomy.

Division group 1	Division group 2
21 - Manufacture of pharmaceutical products	72 - Scientific R&D

Division group 1	Division group 2
33 - Repair and installation of machinery and equipment	77 - Rental and leasing activities
33 - Repair and installation of machinery and equipment	95 - Repair of Computers and personal and household goods
81 - Services to buildings and landscape activites	97 - Activities of households as employers of domestic personnel
59 - Motion picture, video and television programme production...	90 - Creative arts and entertainment activities

The fact that the SIC taxonomy does not capture these relationships is more a limitation of imposing a single hierarchy (in the form of a taxonomy) than a limitation of the SIC taxonomy itself.

Finally, we note that many sectors possess a high degree of similarity to many (if not most) industries - e.g. *Office administrative, office support, and other business support activities (82)* - as they offer services which apply across industries. It may be the case that these services are offered across industries by one business or that each business may specialise in offering those support activities within a specific industry.

4.2 SIC heterogeneity

By aggregating the topic distributions by SIC, and calculating the entropy of the topic distributions for each SIC, we create a measure of the "heterogeneity" of sectors within the Glass data. Figure 9 shows the ten most heterogeneous (highest entropy) and ten least heterogeneous SIC classes according to this measure. The most heterogeneous sectors such as *Other personal services not elsewhere classified* are sectors that a company may be labelled as because their activity is not well-captured by an existing SIC code. Such sectors are prime candidates for reclassification in some way such as adding further levels of depth to the SIC taxonomy for these highly heterogeneous codes - we explore a strategy to do this in Section 5.

Figure 10 plots the distribution of 'topic activity' for the 20 most heterogeneous sectors. The contribution towards 'topic activity' from each SIC code is expressed in terms of the fraction of the mean activity in each topic across all SIC codes - this highlights more important topics and de-emphasises topics which are common to many or all industries. Thus topics on the left of the figure correspond to the topics which are over-represented in heterogeneous SIC codes, which contain terms around digital marketing, consultancy, property management, recruitment, and finance.

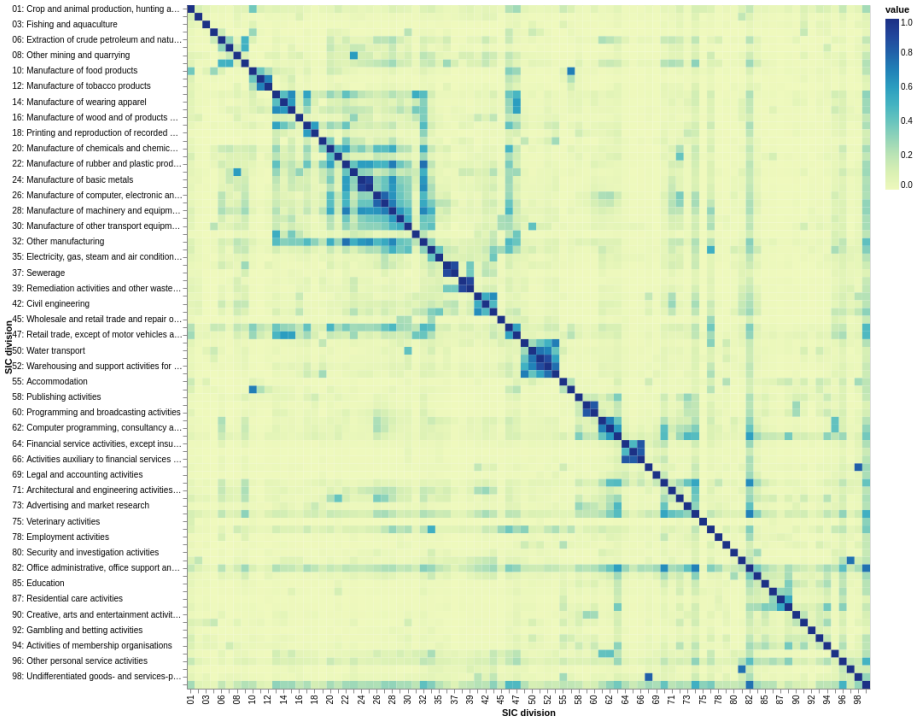

Figure 6: Cosine similarity between SIC divisions similarities (Level 0)

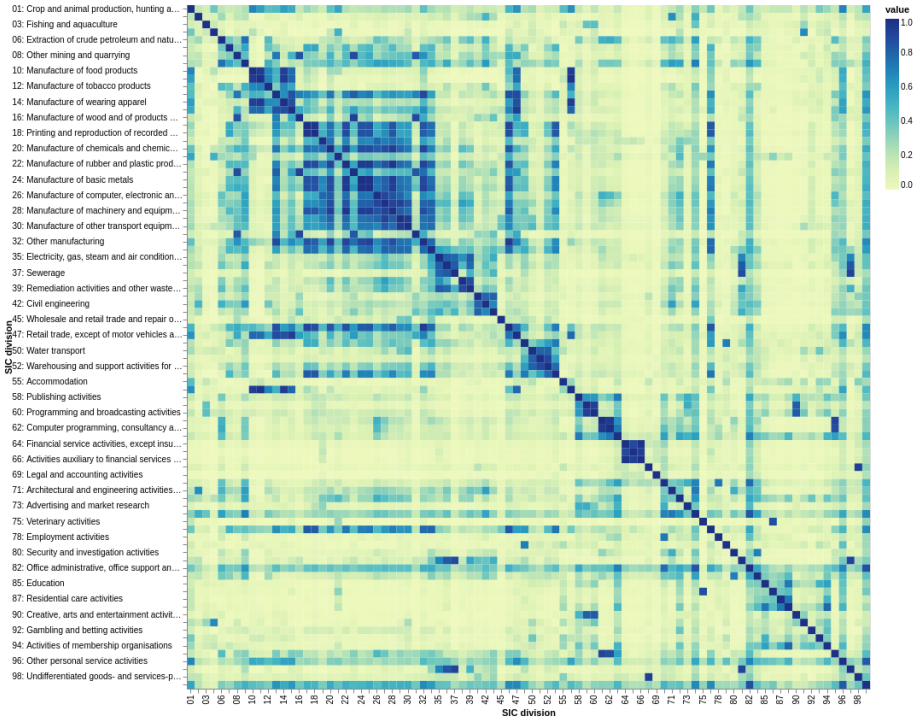

Figure 7: Cosine similarity between SIC divisions similarities (Level 1)

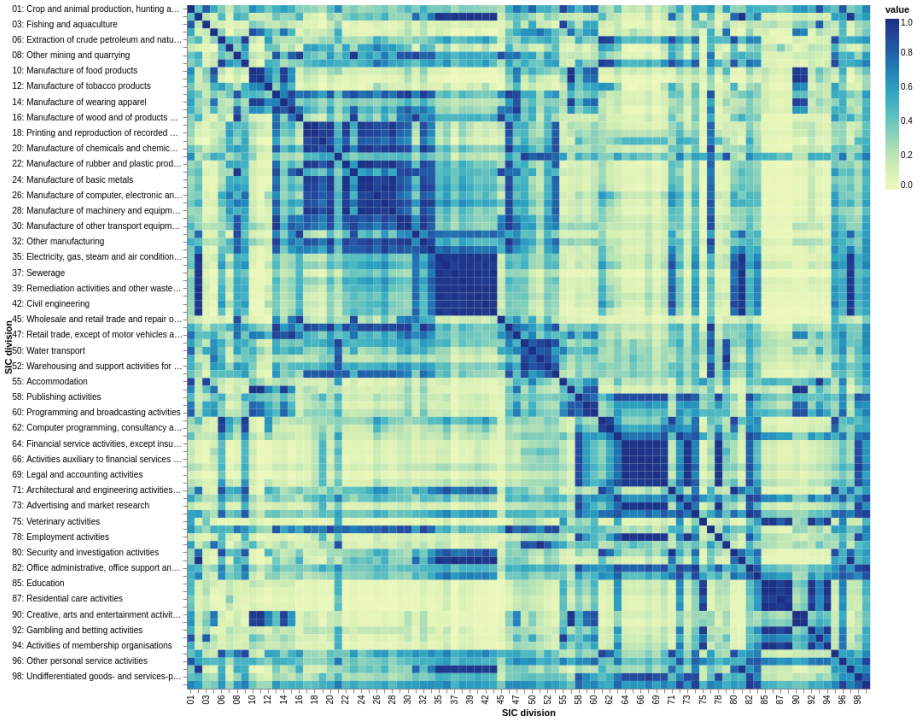

Figure 8: Cosine similarity between SIC divisions similarities (Level 2)

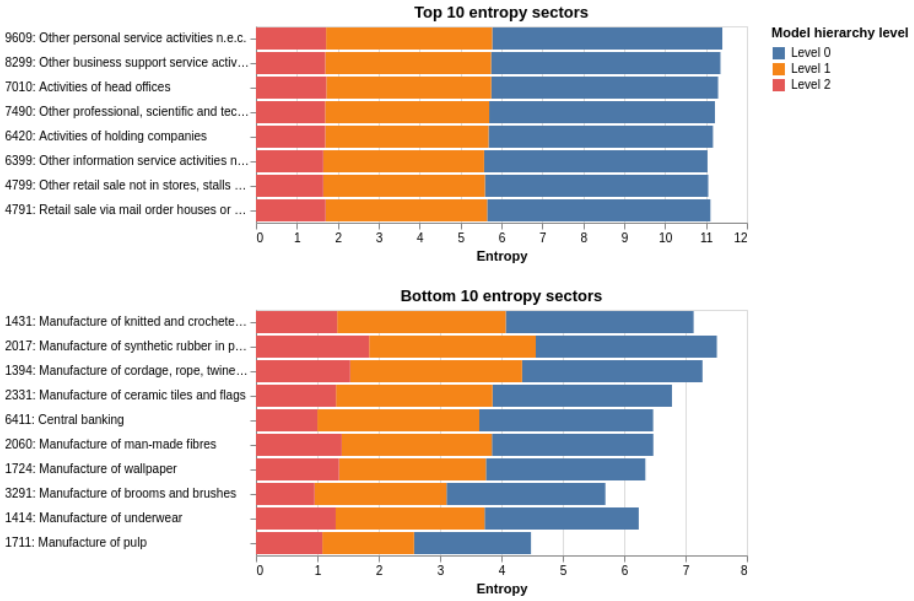

Figure 9: Ten most heterogeneous (highest entropy) and ten least heterogeneous SIC classes calculated according to the entropy of the topic distributions of each SIC class.

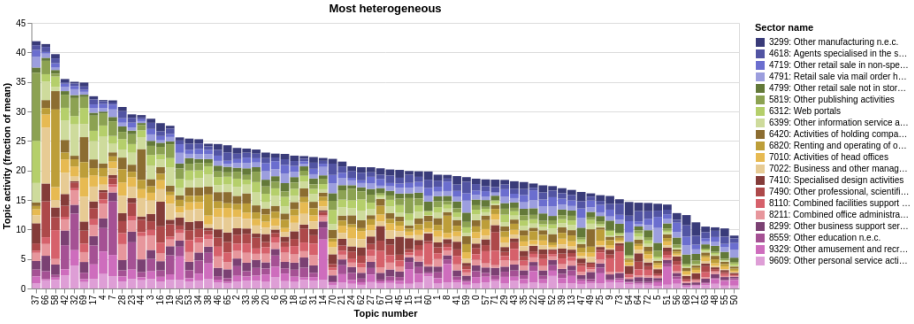

Figure 10: Distribution of topic activity amongst the top 20 most heterogeneous (according to entropy of the topic distributions with SIC codes) sectors. Activity for a topic-sector combination is expressed as a fraction of the mean activity of that topic across all sectors. Based on level 1 of the model hierarchy.

4.3 Conclusion

The analysis of this section has highlighted the rich structure which can be captured by business website description data - picking up both structure defined by the SIC taxonomy and structure that is not captured by the SIC taxonomy.

Furthermore, by analysing the heterogeneity of SIC codes (based on semantic overlap) we can identify the small parts of the SIC taxonomy that are less useful and may benefit from reclassification or a more granular description that captures emerging industries.

The simultaneous construction of a hierarchy of topics and clusters would be a prime candidate for bottom-up taxonomy creation were it not for the fact that the method did not scale to the full Glass dataset. Next section we leverage the structure of the SIC taxonomy to implement a similar clustering strategy in a more scalable way.

5 Pilot strategy to develop a bottom-up industrial taxonomy

Sections 3 and 4 have provided empirical evidence of important limitations of the SIC-2007 taxonomy. They include a mismatch between business descriptions and codes for some SIC codes, a strong overlap between codes in different parts of the taxonomy, and a presence of semantically heterogeneous 'not elsewhere classified' codes that attract very different types of companies as well as some companies that could be classified in other sectors of the taxonomy.

In this section we pilot an approch to build a bottom-up taxonomy based on company descriptions that addresses some of these limitations. Here are some of its features:

1. It draws on the SIC-2007 taxonomy: instead of replacing SIC-2007 with a new structure, we start at a relatively low level of resolution (SIC-4) in the SIC-2007 hierarchy and seek to decompose those categories into more granular sectors through a network analysis that we detail below.[8]
2. Its production is automated: The categories within the taxonomy are identified and named automatically with some manual tuning of specific parametres to increase the intepretability and relevance of outputs.
3. It can be used to tag companies with labels for multiple sectors: This helps overcome the rigidities of the SIC taxonomy by accommodating companies that operate in several industries. it also makes it possible to remove 'duplicate sectors' extracted from different SIC codes, and, potentially to

[8] We note that there is a five-digit SIC category below the four digit but this includes a small number of codes - 191 versus 650 in the SIC4 - that tend to add limited information to what is available at the higher level.

reconstruct the hierarchy of the bottom-up industrial taxonomy through an agglomerative clustering of the network of sector co-occurrences in companies.

In order to pursue these goals, we build a complex pipeline that we describe in the rest of this section, presenting emerging findings as we go. Before doing that, we provide a high-level methodological narrative that summarises the conceptual model informing our approach and the steps we have taken to implement it (see figure 11).

5.1 Methodological narrative

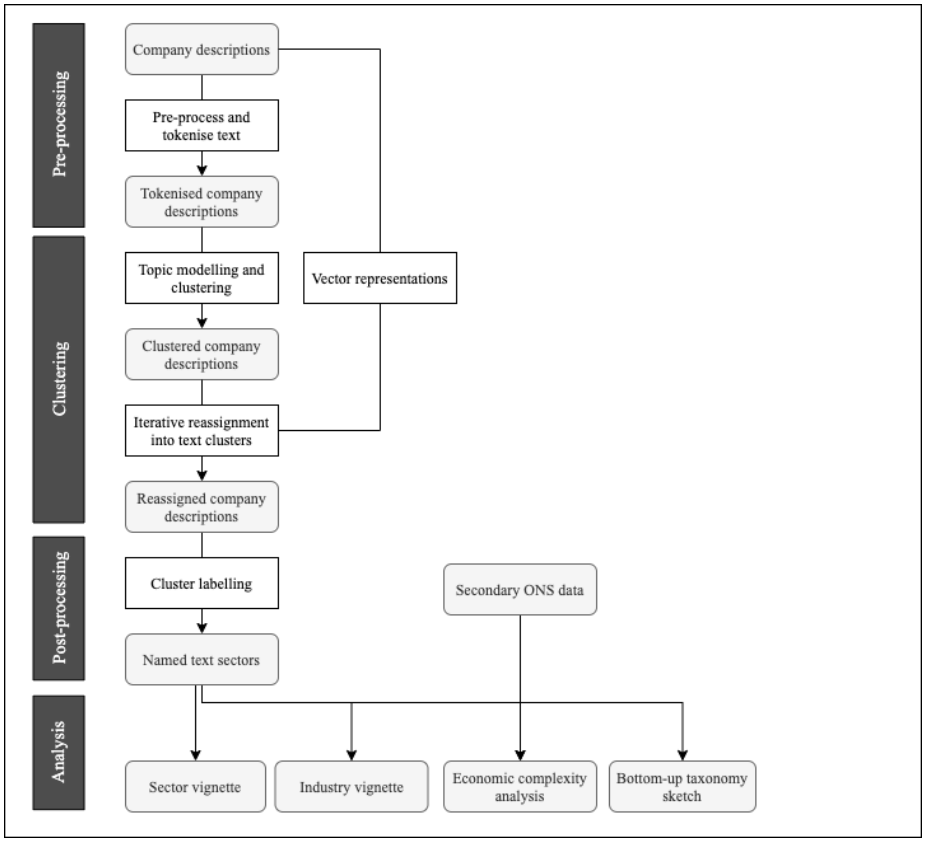

Figure 11: Summary of our pre-processing, analysis and post-processing pipeline

We think of industrial sectors as latent constructs that 'generate' words which are distributed over company descriptions.[9] Our analytical challenge is to infer

[9] We use the term 'word' loosely to include unigrams, bigrams, trigrams etc.

sectors from those word distributions in the Glass data. To do this, we implement a network-based topic modelling strategy that assumes that groups of words that regularly co-occur in company descriptions are manifestations of sectors. The approach we follow, which models co-occurrences of words in company descriptions and similarity of company descriptions based on the words that co-occurr in them symmetrically, allows us to cluster companies based on their text descriptions. We consider these clusters as proxies for industries, and refer to them as 'text sectors.'

This clustering is initially implemented *inside* 4-digit SIC codes for reasons of scalability, precision and granularity. The downside is that we will also misclassify companies that were originally in erroneous SIC4 code. To address this, we use the vector representation of companies created in Section 3 with the DistilBERT model to iteratively reassign companies to the text sectors *across the SIC taxonomy* that they are semantically closest to. An added benefit of this reassignment is that we would expect uninformative / noisy text sectors to lose companies in favour of more informative ones.

Having reassigned companies to their closest text sector, we label those text sectors with the most salient terms in their corpora, and perform some exploratory analyses of the results by:

- Examining the composition of SIC 7490, one of the highly heterogeneous 4-digit SIC codes we identified above.
- Examing "green economy" activities across the population of text sectors
- Benchmarking metrics of the economic complexity of local economies in the UK based on SIC4 and text sectors.

Our sector reassignment step also yields a list of the K-closest sectors to a company that can be used for multi-sector tagging and to build text-sector similarity networks which can be clustered hierarchically to build an industrial taxonomy - we conclude with an experimental visualisation of such a taxonomy.

5.2 Text preprocessing

We start our analysis with just over 340,000 business website descriptions that have been matched with Companies House with a high level of certainty (match score above 75). The dataset includes 611 SIC4 codes (99% of all SIC4s). We have also tokenised their descriptions, performed named entity recognition to identify tokens that represent locations, company names, times etc. and removed those, and estimated n-grams (i.e. combined pairs and triads of tokens that occur frequently in the corpus, such as `city` and `centre` into a single token `city_centre`).

5.3 Topic modeling

We want to cluster companies within 4-digit SIC codes into a more granular set of categories. To do this, we will use `top-SBM`, the same topic modelling algorithm based on the stochastic blockmodel that we used in Section 4 during our analysis of SIC4 heterogeneity (**tobSBM?**). TopSBM identifies topics in the corpus by building a network of word cooccurrences in documents and partitioning it into the set of communities (topics) most likely to have generated the data. This same approach can be used to represent documents in a network connecting those documents that share words (topics). This approach is applied recursively to generate a hierarchy of topics (document clusters) in the data, which are identified automatically. Here, we will focus our analysis on the outputs of the document (company description) clustering, which we treat as 'seed text sectors' for subsequent stages of the analysis.

Our decision to perform the document clustering *inside* 4-digit SICs has four reasons:

1. As noted in the introduction to the section, we are interested in adding an extra level of granularity to the existing SIC taxonomy. Using the existing company classification as the first step for additional decompositions is an obvious strategy to achieve this.
2. The `topSBM` algorithm that we use has several advantages over alternative topic models and / or unsupervised machine learning approaches (e.g. clustering) that we could use to segment documents but it has the downside that it is harder to scale to larger corpora. By decomposing the corpus into a smaller set of 4-digit SIC sub-corpora we are able to deploy `topSBM` more efficiently.
3. Even though, as noted before, 4-digit SIC codes are noisy, our prior is that they still contain useful information for classifying companies into sectors. In particular, we would assume that companies in the same 4-digit SIC code will interpret words in similar ways, reducing the scope for semantic ambiguity that may introduce noise into our topic modeling if we performed it at the level of the whole corpus (for example, the term "network" will have different meanings in telecommunications, professional services and health-related SIC codes - if we implement our topic model in a corpus that incorporates all these sectors, network may be classified into an uninformative topic comprising the disparate terms it co-occurs with in each of them).
4. Performing our analysis inside 4-digit SIC codes increases the granularity of the decomposition we are able to perform and reduces the risk that particular sectors dominate the clustering.

This approach is not without limitations - we highlight them and how we have addressed them in section 5.4.

We focus our topic modelling on 43 4-digit SIC codes with more than 2,000

companies comprising just over 192,000 companies. The topic model and company description clustering yields 1,884 text sectors. In figure 12 we present, for each of the SICs where we have clustered company descriptions, the number of text sectors extracted (size of the ciccles) and the minimum description length of the topic model trained in the corpus (a measure of the amount of information required to describe the corpus, providing an indicator of the goodness of fit of the topic model).

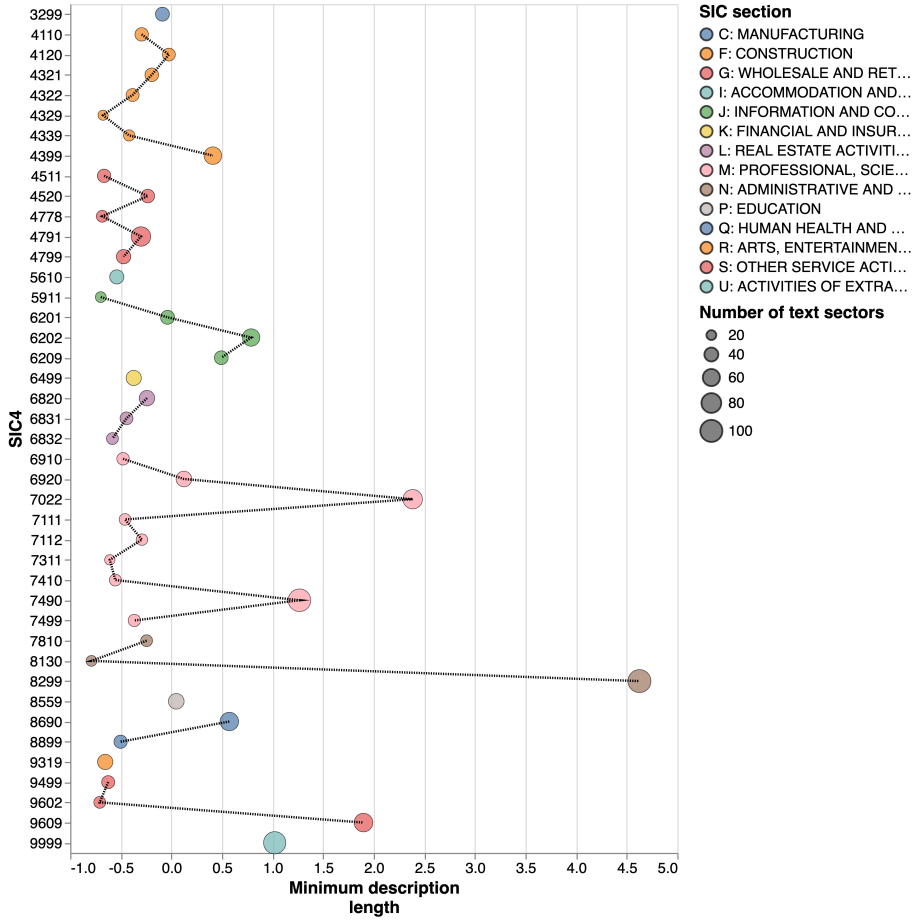

Figure 12: Topic model / document clustering outputs by 4-digit SIC. The size of the circles represents the number of text sectors extracted from the SIC, and the horizontal axis the minimum description length of the topic model trained on the sector.

We note that "Not Elsewhere Classified" SIC codes such as 8299 (Other business support service activities), 9609 (Other personal service activities) and 7490 (Other professional, scientific and technical activities) yield a large number of

sectors and require more information to be described by the model, consistent with the idea that they are particularly fruitful sites for additional sector decomposition. Other sectors with large text sector populations and minimum description lengths include wide-ranging activities such as 7022 ("Business and other management consultancy activities") and 6022 ("Computer consultancy activities"). On the other hand, some of the sectors that yield few communities and are easy to describe by our topic models include 4329 ("Other construction installation"), 5911 ("Motion picture, video and television programme production activities") or 9602 ("Hairdressing and other beauty treatment") - these are arguably narrower sectors with much more informative SIC descriptions where we would expect to find a more homogeneous set of companies.

5.4 Sector reassignment

The main downside of our clustering strategy is its restrictiveness: companies are clustered in groups inside 4-digit SIC codes. This means that a company that had been initially misclassified in a SIC code will be misclassified in a text sector (although we note that if companies from an industry are *systematically misclassified* into a 4-digit SIC, our analysis may be able to identify them as a cluster within that SIC). In addition to this, companies in new industries for which there is not currently a SIC code will be scattered across text sectors in different SICs (e.g. in assorted "not elsewhere sectors") instead of being clustered together.

In order to address this, we implement a recursive sector reassignment step that draws on the vector representations of company descriptions that we obtained in section 3. This involves the following steps:

1. We calculate, for each of the text sectors in our data, a semantic centroid which is the mean of the vector representations of all the companies assigned to that text sector initially. One way to think about it is as the 'signature' or semantic summary of that text sector.
2. We calculate, for all companies in our data, their L1 (Manhattan) distance to each of those centroids. We do this efficiently using FAISS, a tool for rapid similarity search with high-dimensional vectors (Johnson, Douze, and Jégou 2019).
3. We reassign each company to the text sector that it is closest to
4. We repeat steps 1-3 over 14 iterations.

figure **??** shows the share of companies changing text sector / SIC code (in other words, being reassigned from a text sector belonging to one SIC to a text sector belonging to another) or changing SIC division (i.e. 2-digit SIC code). After an initial stage of turbulence where less than 20% of companies are reassigned to their original text sector (justifying the adoption of this strategy), we see the assignments stabilise. By our last iteration a negligible number of companies are still being reassigned between text sectors.

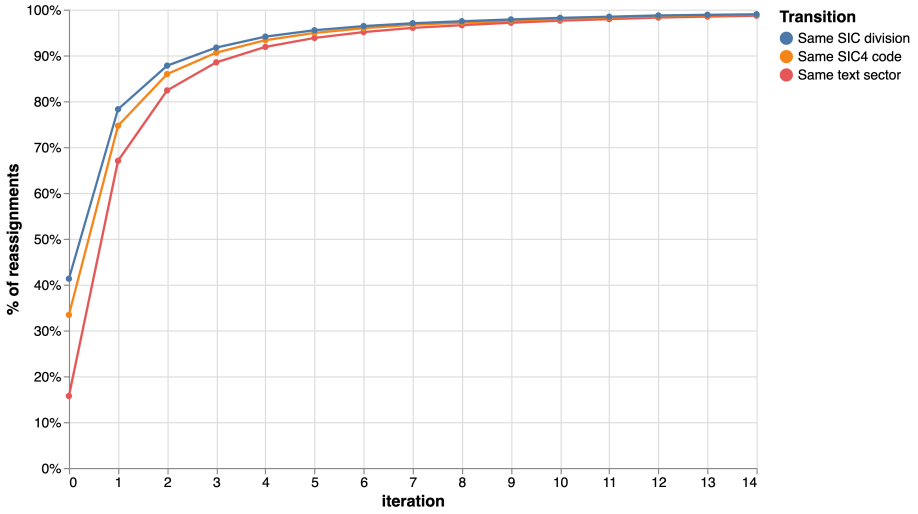

Figure 13: Share of companies reassigned to the same text sector / 4-digit SIC code / SIC division at each iteration of the reassignment process

In figure 14 we present the origin and destination of sector transitions during the reassignment procedure, both in terms of individual text sectors and of the 4-digit SIC codes for text sectors. In both cases we see more activity in and around the diagonal, consistent with the idea that in a majority of cases companies stayed in their original text sector / SIC code. This suggests that the original clustering procedure generated many informative and robust clusters. At the same time, we see substantial transition levels betweem sectors outside the diagonal. For example, 10% of companies initially in text sectors within SIC4 8130 (Landscape Service activities) transition to SIC4 4399 (Other contruction activities). 7% of companies initially in text sectors belonging to SIC 6499 (Other financial service sector activities) transition to text sectors in SIC 6920 (Accounting and book-keeping). We also note substantial company flows involving "other" and "not elsewhere classified" text sectors.

What happens to "low quality" text sectors during the reassignment process? Our prior is that fuzzy and imprecise text sectors might lose companies to more accurate ones (e.g. because their centroid is, on average, further away from the vector representations of companies in the sector). We explore this hypothesis by comparing, for each text sector, its reassignment ratio (number of companies 'won' over number of companies lost during the reassignment process) with its heterogeneity, which we measure as the average distance of a vector representation of each company in the text sector to its centroid. If our prior is correct, we would expect more heterogeneous text sectors to lose companies and more homogeneous text sectors to win companies.

We present the results in figure 16. Our analysis reveals a weak negative corre-

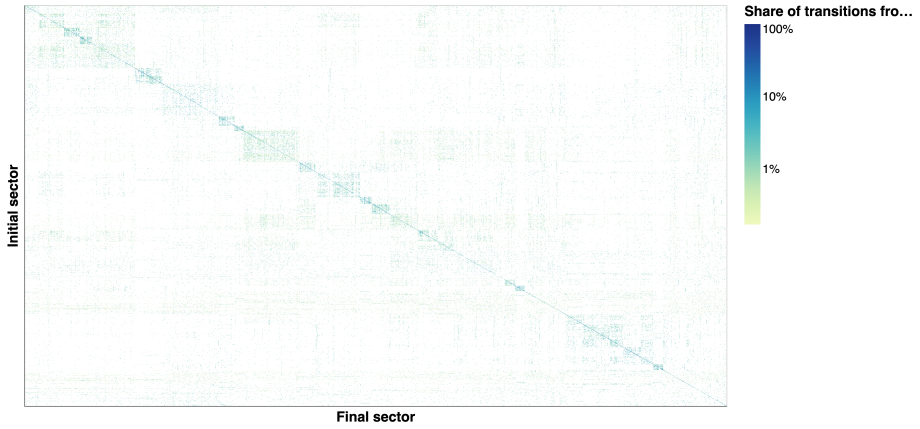

Figure 14: Share of transitions from an origin text sector (vertical axis) to a final text sector (horizontal axis) after 15 iterations of the sector reassignment procedure

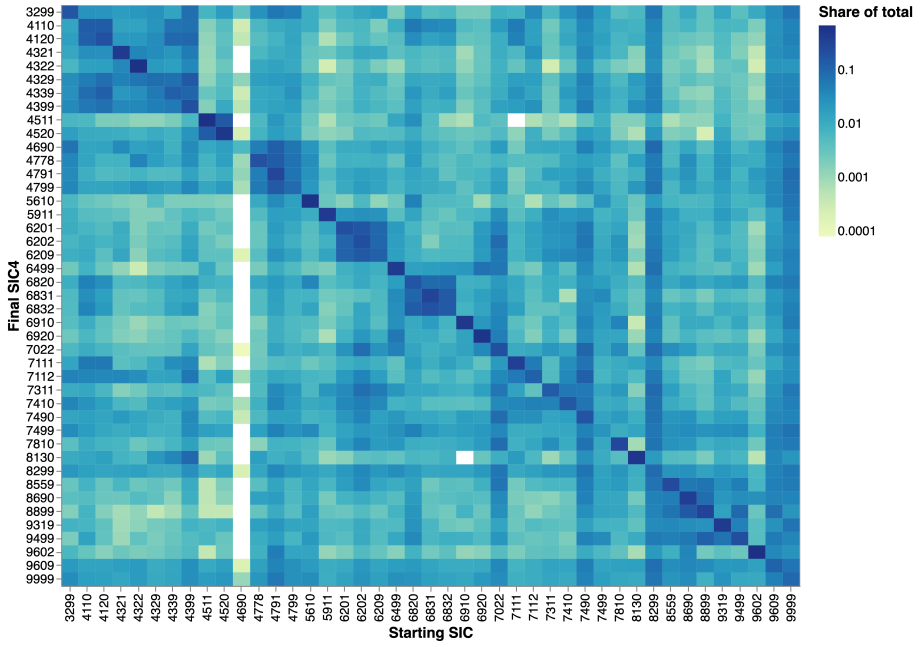

Figure 15: Share of transitions from a text sector in an origin 4-digit SIC code (vertical axis) to a text sector in a final SIC code (horizontal axis) after 15 iterations of the sector reassignment procedure

lation between text sector heterogeneity and its reassignment ratio (Spearman $\rho = -0.31$), consistent with what we expected, and suggesting that our reassignment process penalises lower quality, less consistent sectors. We note that services text sectors (yellow and blue circles tend to more heterogeneous than manufacturing and construction, perhaps suggesting additional scope for decomposition into even more granular text sectors).

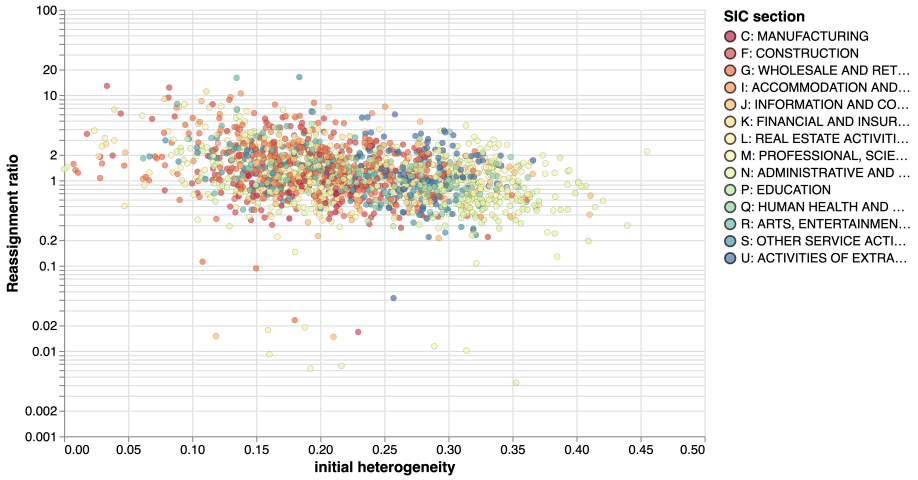

Figure 16: Comparison between a text sector's heterogeneity and its reassignment ratio during the reassigment process (values over one represent situations where a text sector wins companies during the reassignment process). The colour represents the SIC section that the text sector's SIC belongs to.

5.5 Postprocessing

Having reassigned companies between text sectors, we proceed to name them by concatenating all company descriptions inside each text sector and identifying the most salient (highest term-frequency inverse-document-frequency (TF-IDF)) words in them.

We present twenty randomly drawn examples in table **??**. These examples suggest that our approach is identifying meaningful sectors in previously uninformative categories (e.g. children related product and services in SIC 9999 (Unclassifiable) or vaping products in 4778 (other retail activities)) and decomposing aggregated sectors into more finely grained activities that might be of interest (e.g corporate social responsibility activities within management consultancy, digital and social media marketing in Specialised Design Activities).

Table 6: Randomly selected example text sectors {#tbl:4.1}

SIC4	Text sector ID	salient_terms
9999: Unclassifiable	0	child baby toy parent kid school session activity bedroom childcare
9999: Unclassifiable	82	repair physiotherapy removal accident_repair refurbish wheel refrigeration mot_testing mezzanine_floor machine_tool
6202: Computer consultancy activities	45	software software_development platform application mobile integration cloud integrate open_source consultancy
4778: Other retail sale of new goods in specialised stores	16	e_liquid flavour vape nut liquid bottle juice price fresh accessory
4791: Retail sale via mail order houses or via Internet	56	garden plant gardening basket seed vegetable flower grower food nursery
6201: Computer programming activities	6	tech love innovation digital positively_impact idea passionate gaming geek thing
4799: Other retail sale not in stores, stalls or markets	28	collector item auction memorabilia watch stock rare antique collection gun
8899: Other social work activities without accommodation n.e.c.	4	activity centre child club play volunteer resident young_people school artist
4399: Other specialised construction activities n.e.c.	54	pickup contractor engineer repair contract roller_shutter maintenance garage maintain domestic_commercial
7410: Specialised design activities	0	collection print interior contemporary textile colour art rug homeware fabric
8690: Other human health activities	16	art young_people child_young mental_health tape counselling child activity self_esteem creative
9499: Activities of other membership organisations n.e.c.	10	tile grant closure committee trustees adam canal registered_charity bowl form
7022: Business and other management consultancy activities	4	corporate_responsibility corporate_social sustainability responsibility employee corporate_governance impact stakeholder activity csr
3299: Other manufacturing n.e.c.	7	packaging print printing label design_print large_format colour marquee printer clothing
7410: Specialised design activities	23	marketing website creative strategy social_medium web_design digital_marketing campaign success digital

SIC4	Text sector ID	salient_terms
4520: Maintenance and repair of motor vehicles	0	garage repair diagnostic_equipment diagnostic ecu workshop fault car_servicing mot make_model
6499: Other financial service activities, except insurance and pension funding, n.e.c.	34	financial finance growth platform bank software lack marketing success economy
8690: Other human health activities	68	foot treatment clinic patient surgery nail treat condition health pain_free
4321: Electrical installation	24	energy maintenance network contractor sustainable safety engineering engineer cost_effective utility
8690: Other human health activities	14	charity donation donate raise_money fund fundraising volunteer hospice raise_fund school

5.6 Exploratory analyses

This subsection explores in further detail some of the opportunities opened up by our prototype text-based industrial classification.

5.6.1 Unpacking SIC 7490

We begin with an exploration or the results of our sector decomposition, reassignment and labelling in SIC 7490 ("Other professional, scientific and technical activities N.E.C."). This is a SIC code which previous analysis in this report showed displaying high levels of heterogeneity. We might also expect to find inside it knowledge intensive companies with high growth potential that operate in 'new sectors' not yet captured by the lagging SIC taxonomy.

figure 17 presents the distribution of companies over text sectors in this SIC. The colours of the bars represent the original SIC sections for text sectors before reassignment. Our results highlight the heterogeneity of sectors contained in 7490, ranging from `7490_29` (coaching) to `7490_50` (mediation), `7490_90` (renewable energies) or `7490_52` (copywriting). This is also reflected in the section origins for companies in text sectors such as `7490_66` (construction) or `7490_57` (health). We also note the presence of some noisy / hard to interpret text sectors in the data such as `7490_5`, which seems to capture companies that mention the positions and names of individual employees in their website descriptions. This points at the need to implement strategies to remove spurious text sectors in future iterations of the analysis.

5.6.2 Green economy text sectors in the Glass data

We can also identify text sectors in policy-relevant areas more effectively than in the SIC taxonomy. As an illustration, we have selected all text sectors that mention three

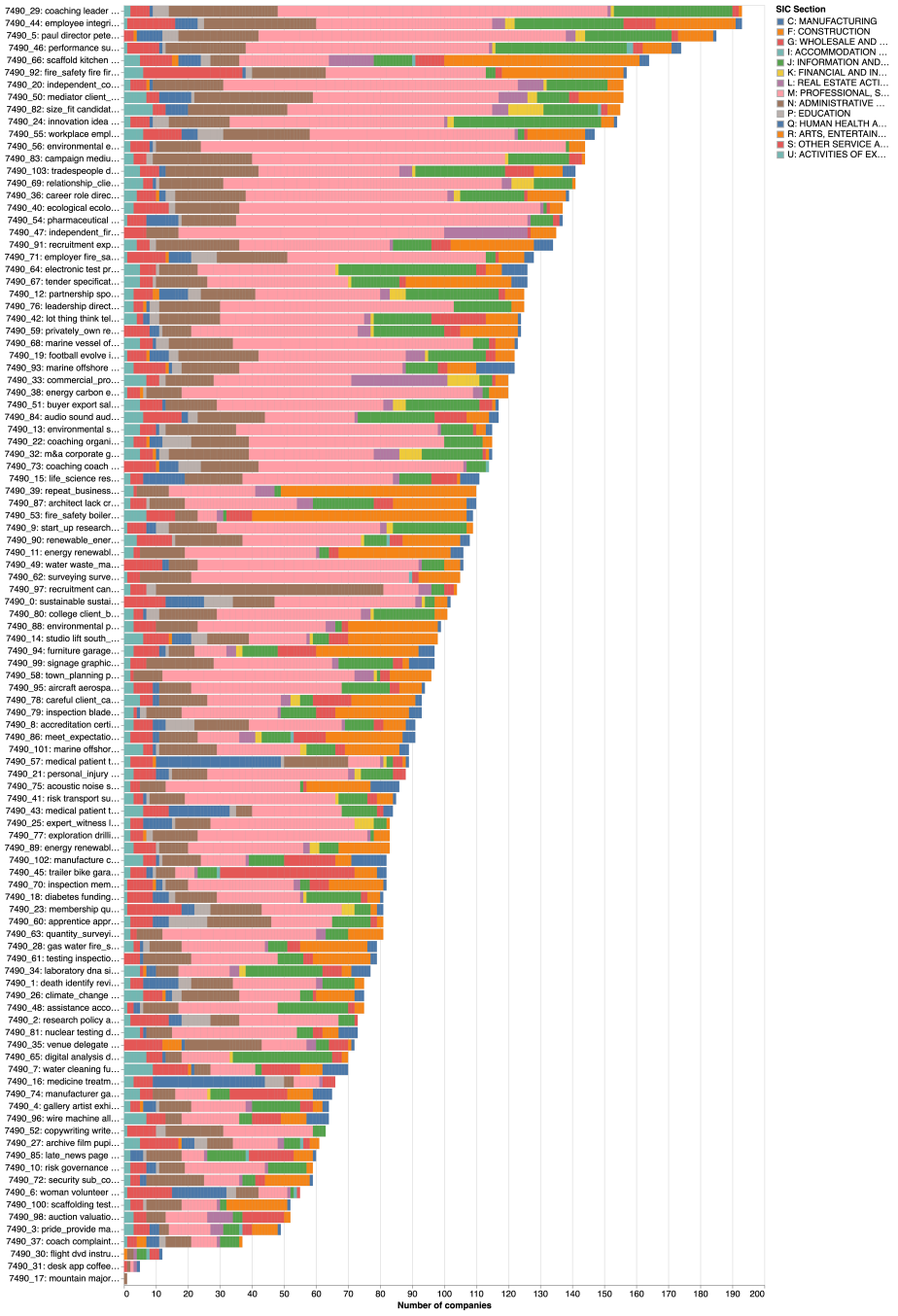

Figure 17: Distribution of companies over text sectors in SIC 7490. The colour represents the SIC section of a origin text sector before reassignment.

or more terms related to the green economy such as “environmental,” “renewable,” “solar,” “sustainability,” “energy,” “emission,” “sustainable,” “green_energy” in their labels. We identify 25 sectors comprising 2,508 firms.

We present their distribution by text sector and source sector in figure 18. The chart illustrates the range of economic activities related to the green economy that we are able to identify in our data, ranging from sustainable housing (`4110_1`) to solar panels (`4321_23`) and wind turbines (`4799_27`), thermal energy (`7112_2`), waste management (`8299_7`) and environmental litigation (`9609_28`).

There is also some repetition in the labels for the sectors we have identified. This could be an artifact of our strategy to label sectors (i.e. we extract each text sector’s salient terms compared to all other text sectors in the corpus, potentially creating a situation where several text sectors share the same terms in their labels). It also suggests that we might be able to merge similar text sectors extracted from different SIC codes, an idea we explore further below.

Figure 18: Distribution of companies over text sectors in 25 sectors mentioning environmental terms in their labels. The colour represents the SIC section of a source text sector before being reassigned to a text sector.

5.6.3 Economic complexity benchmarking

Here, we use our text-sector classification to produce estimates of economic complexity ("Economic Complexity Indices" or "ECI") at the local authority level in the UK and estimate the association between these indices and other measures of local economic performance and knowledge intensity such as GDP per capita, GDP growth, median annual earnings and share of the workforce with tertiary education.[10] We then compare the strength of those associations with an ECI based on 4-digit SIC codes.

This analysis is motivated by strong evidence of a link between ECI - a composite indicator capturing the relative uniqueness of the productive capabilities in a location - and other metrics of local economic performance Bishop and Mateos-Garcia (2019). If text sectors offer a more accurate representation of the productive capabilities in a location than aggregated and -in some cases- uninformative SIC codes, then we would expect an ECI based on it to be more strongly associated with economic outcomes in a location.

We explore this hypothesis using a Bayesian linear regression framework implemented in the BAMBI package Gelman, Hill, and Vehtari (2020) where we regress our measures of economic performance based on secondary data on ECI indices calculated with text sectors and 4-digit SIC codes controlling for local population (logged) and a local authority's NUTS-1 region to account for potential economies of scale and systematic differences in the economic performance of different UK regions. We present the 94% high density intervals in figure 19.

All local economic performance and knowledge intensity indices show a stronger association with the text-based ECI than with the SIC-based ECI. When we compare pairs of models using leave-one-out cross-validation (Vehtari, Gelman, and Gabry 2017) we find that, in all cases, the models based on text-based ECI indices outperform the SIC-based ECIs. This supports the idea that our text classification can help characterise the economic composition of a location (and perhaps its strengths and weaknesses) more accurately than the SIC alternative.

5.7 A hierarchical taxonomy

We conclude our exploration of results by building a hierarchical industrial taxonomy based on the semantic similarity between text sectors. In order to do this, we extract the five closest sectors to each company at the end of the sector reassignment strategy described before and use it to build a text sector co-occurrence network connecting those text sectors that are closer to each other.[11]

Having done this, we decompose the network into a hierarchical set of densely-connected communities using a stochastic block-model (Peixoto 2017) - this results in a network where an initial layer of 1,884 text sectors is connected to progressively more aggregated communities. We display a network with three levels of the hierarchy above that lowest (text sector) level.

[10] We extract these secondary data from ONS regional accounts data, the Annual Survey of Hours and Earnings and the Annual Population Survey.

[11] The closest sector to each company would be the one it was classified into at the end of our sector reassignment procedure.

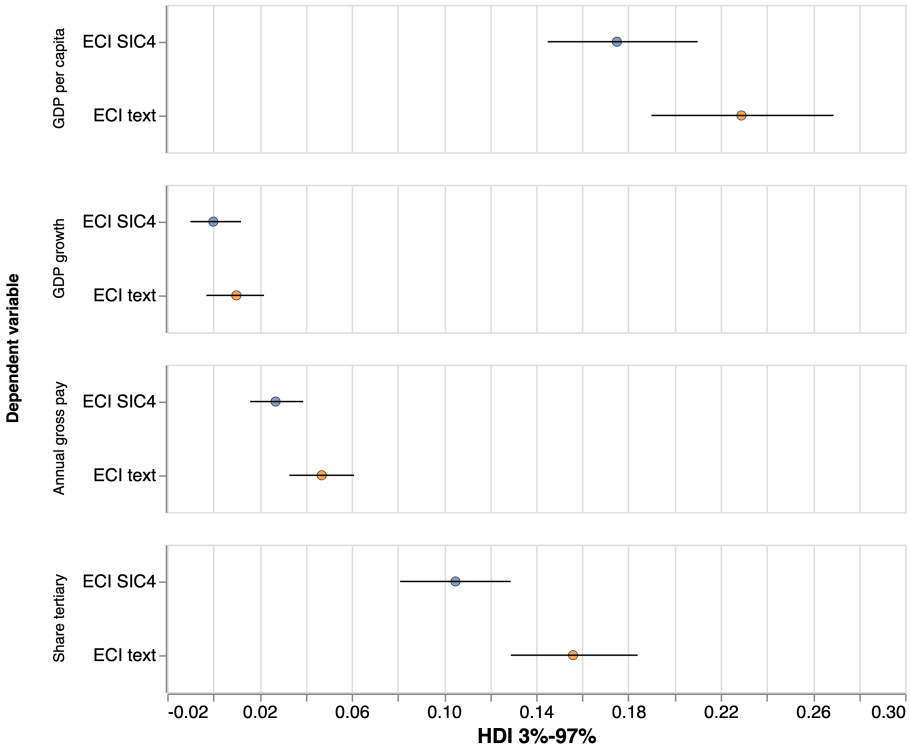

Figure 19: High Density intervals for estimates of the association between ECI indices based on text sector classification and SIC sector classification and various measures of local economic performance.

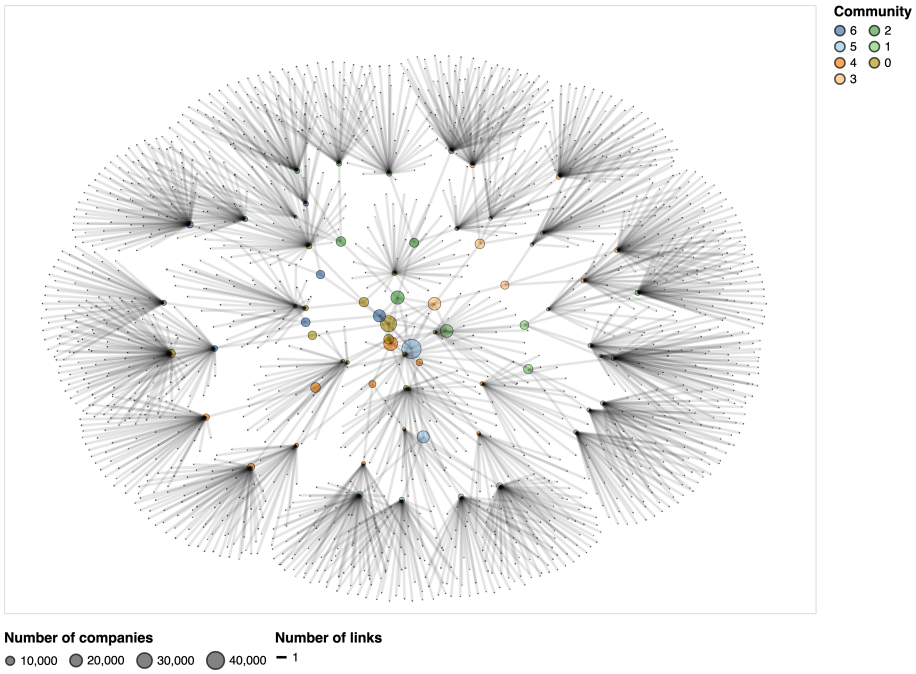

Figure 20: Prototype hierarchical industrial taxonomy

At this point, this taxonomy is just the experimental output of one strategy to transform a 'flat' collection of text sectors into a hierarchy that can be explored and aggregated at various levels of resolution. A future step in the analysis will be to analyse its interpretability, its comparability with the SIC taxonomy and the role that it can play in deduplicating similar text sectors extracted from different parts of the SIC taxonomy.

5.8 Conclusions and next steps

The pipeline we have developed yields a highly granular collection of text sectors that can be used to decompose noisy and uninformative SIC codes, explore policy relevant industrial activities such as those related to the 'green economy,' and capture the industrial composition of local economies more accurately than is possible with the SIC taxonomy. There are many other possibilities that we have not explored, such as for using text sectors to proxy the presence of particular capabilities in a company, such as for example the adoption of digital and data analytics technologies.

Our analysis is, however, not without limitations that we will seek to address in future work.

1. Our pipeline only includes 43 4-digit SIC codes. Future analysis should seek to improve our coverage of sectors in the Glass data.
2. The text sector extraction procedure yields some noisy sectors based on irrelevant content in company descriptions. Going forward, we will explore avenues to reduce their presence through text-preprocessing (e.g. removing potentially irrelevant text from company descriptions before training our topic models) or post-processing (e.g. analysing various sector features such as their homogeneity and removing them before the sector reassignment stage).
3. Our procedure to reasign companies to text sectors is binary: each company is assigned to its closest sector regardless of its position within the distribution of distances to that sector. It would be preferrable if this procedure generated a measure of confidence in the assignment which could then use to test the robustness of our results to changes in similarity thresholds.
4. The sector labelling does not yield sufficiently distinctive text sector names. This could be addressed by calculating TF-IDF scores for a sector's corpus compared to similar text sectors (for example those extracted from the same SIC code) instead of the whole corpus.
5. We noted the presence of duplicated text sectors extracted from different SIC codes in the data. We will explore strategies to merge duplicated text sectors, for example, by implementing an agglomerative clustering step along the lines of what we have done to build the prototype hierarchical taxonomy.

6 Conclusion

Our analysis of a labelled dataset of SIC codes and company descriptions obtained from business websites has helped to assess the limitations of the current SIC taxonomy, identify opportunities to improve it using natural language processing and network analysis and some of the new challenges created by this approach. Here we summarise

key issues, discuss next steps for the research and highlight broader implementation factors.

6.1 Limitations of the current taxonomy

Our analysis of the SIC taxonomy using supervised and unsupervised methods in Sections 3 and 4 highlights important mismatches between the language that businesses use to describe what they do and the SIC codes where they are classified. More specifically, we find that businesses with similar descriptions are sometimes placed in different SIC codes, and that businesses with different descriptions are sometimes placed in the same codes. Our decomposition of sectors into highly grained communities illustrates the degree of heterogeneity in SIC codes such as 7490, which appears to include companies providing support services to the pharmaceutical industry, companies working with renewable energies and specialist lawyers.

The resulting misclassification is likely to introduce noise into economic indicators about the composition of the economy specially when those are produced at a high level of industrial resolution. While we are aware that some of the sectoral misclassification present in Companies House is likely to be addressed 'downstream' as additional data is collected from businesses, it is unclear how this might help to address situations where businesses are active in areas currently absent from the SIC taxonomy, or where there is ambiguity about their industrial focus because they operate in multiple sectors. The bottom-up taxonomy that we have piloted in Section 5 sets out to overcome some of these limitations.

6.2 Advantages of a new, bottom-up taxonomy

Section 5 illustrates some of the advantages of a new taxonomy based on semantic clustering of companies based on their text descriptions: new ecomomic activities - for example around the green economy - can be detected and studied, and companies can be labelled with multiple sectors thus capturing their diversification or the adoption of particular technologies and practices. We are able to 'open up' the black box of "not elsewhere classified" SIC codes and analyse their composition.

A bottom-up industrial taxonomy may also make it possible to look at the economy from new and potentially useful perspectives e.g. sets of companies that are part of the same value chain or have adopted similar technologies or production processes complementing those offered by the SIC taxonomy.

6.3 Challenges for developing a new taxonomy

Section 5 also shows the challenges for developing a bottom-up industrial taxonomy: this involves a complex pipeline with multiple steps and some hard to interpret results. Our decision to focus on companies with a stronger presence in the Glass data leads to the removal of SIC codes in primary sectors that are less well-covered there. These limitations point at opportunities for improvement and further development that we will pursue in the next phase of the project.

6.4 Next steps

As noted, our next step is to simplify and improve our text processing, sector identification and company classification pipeline, and to use the results in an applied analysis that demonstrates the value added of the taxonomy. Some options for such analysis include:

- Analysing the economic geography and performance of a new and policy relevant sectors such as for example the green economy, that are currently absent from the SIC taxonomy.
- Analysing growth dynamics in bottom-up sectors of interest after matching a firm-level dataset labelled with new sectors and micro surveys such as IDBR, ABS or the Community Innovation Survey
- Implementing an alternative, hierarchical bottom-up taxonomy and comparing its geography and evolution with SIC-2007. This will require developing methods that draw on official, comprehensive micro-data to estimate indicators of economic performance such as employment in new sectors.

6.5 Implementation considerations

We conclude by highlighting additional considerations for the implementation of a bottom-up industrial taxonomy based on company descriptions fron their websites:

1. Coverage: As we previously pointed out, business websites may fail to cover some industries, which could create gaps in the taxonomy. One way to address this would be to deploy the bottom up industrial taxonomy as a tool to improve the resolution and granularity of analysis of knowledge intensive activities which our analysis suggest are particularly poorly-served by the industrial taxonomy, while preserving existing codes for other sectors.
2. Estimating employment levels: Business websites offer rich information about the markets that a business serves but miss important dimensions of economic activity such as employment or productivity. One way to address this gap is by matching company descriptions and sectors with official micro-data such as for example IDBR or ABS drawing on Companies House numbers we have obtained through the fuzzy matching protocol outlined in Section 2. An additional step would be required to estimate population level statistics in various sectors from the incomplete sample of companies we have access to.
3. Longitudinal updates: A bottom-up industrial taxonomy could in principle be updated close to real-time as the economy evolves and new industries appear and are reflected in company descriptions. While this could offer a very timely perspective on the composition of the economy, it would come at the cost of longitudinal consistency in terms of our ability to study the evolution of industries over time. One potential strategy to manage this trade-off would be to maintain a frequently updated bottom-up industrial taxonomy at the sub-SIC4 level with less regular updates above that. One advantage of this approach is that it would help to identify nascent sectors with a critical mass of activity that might warrant the creation of higher-level codes.

These considerations suggest that in order to benefit from novel data sources and methods, the economic statistics system will need to innovate in the infrastructures it uses to connect open and web sources with official microdata, to monitor the evolution of the economy in order to identify the emergence of new sectors close to real time, and to adopt machine learning methods in order to transform that enhanced understanding into up-to-data industrial taxonomies offering the granular and timely views of the economy and its constituent industries that policymakers increasingly demand.

Bibliography

Bean, Charles R. 2016. *Independent Review of UK Economic Statistics*. HM Treasury.

Bishop, Alex, and Juan Mateos-Garcia. 2019. "Exploring the Link Between Economic Complexity and Emergent Economic Activities." *National Institute Economic Review* 249: R47–58.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3: 993–1022.

Broder, Andrei Z, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. 2000. "Min-Wise Independent Permutations." *Journal of Computer and System Sciences* 60 (3): 630–59.

Capretto, Tomás, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A Martin. 2020. "Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python." *arXiv Preprint arXiv:2012.10754*.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.

Gerlach, Martin, Tiago P Peixoto, and Eduardo G Altmann. 2018. "A Network Approach to Topic Models." *Science Advances* 4 (7): eaaq1360.

Hicks, Diana. 2011. "Structural Change and Industrial Classification." *Structural Change and Economic Dynamics* 22 (2): 93–105.

Hidalgo, César A, and Ricardo Hausmann. 2009. "The Building Blocks of Economic Complexity." *Proceedings of the National Academy of Sciences* 106 (26): 10570–75.

Hughes, John C, Gareth James, Andrew Evans, and Debra Prestwood. 2009. "Implementation of Standard Industrial Classification 2007: December 2009 Update." *Economic & Labour Market Review* 3 (12): 51–55.

Johnson, Jeff, Matthijs Douze, and Herve Jegou. 2019. "Billion-Scale Similarity Search with GPUs." *IEEE Transactions on Big Data*, 1–1. https://doi.org/10.1109/tbdata.2019.2921572.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. "Billion-Scale Similarity Search with Gpus." *IEEE Transactions on Big Data*.

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." In *Soviet Physics Doklady*, 10:707–10. 8. Soviet Union.

Mateos-Garcia, Juan, Hasan Bakhshi, and Mark Lenel. 2014. "A Map of the UK Games Industry." *London: Nesta*.

Mateos-Garcia, Juan, Konstantinos Stathoulopoulos, and Nick Thomas. 2018. "The Immersive Economy in the UK: The Growth of Virtual, Augmented, and Mixed Reality Technologies."

McInnes, Leland, John Healy, and James Melville. 2018. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*.

Nathan, Max, Simon Adderley, Michele Bernini, Rachel Mulhall, and Paulina Ramirez. 2017. "Industrial Clusters in England."

Nathan, Max, and Anna Rosso. 2015. "Mapping Digital Businesses with Big Data: Some Early Findings from the UK." *Research Policy* 44 (9): 1714–33.

Peixoto, Tiago P. 2017. "Nonparametric Bayesian Inference of the Microcanonical Stochastic Block Model." *Physical Review E* 95 (1): 012317.

Rajaraman, Anand, and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge University Press.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *arXiv Preprint arXiv:1910.01108*.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27 (5): 1413–32.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.