# Using text data to build a bottom-up industrial taxonomy for 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 . . . . . 9

**3 Using predictive modelling to analyse the SIC taxonomy** **11**
3.1 Modelling methodology . . . . . 12
3.2 Model performance and analysis . . . . . 16
3.3 Conclusion . . . . . 21

**4 Hierarchical Topic modelling** **22**
4.1 SIC similarities . . . . . 22
4.2 SIC heterogeneity . . . . . 26
4.3 Conclusion . . . . . 27

**5 Pilot strategy to develop a bottom-up industrial taxonomy** **28**
5.1 Methodological narrative . . . . . 29
5.2 Text pre-processing . . . . . 31
5.3 Community extraction . . . . . 33
5.4 Post-processing . . . . . 39
5.5 Sketching a bottom-up industrial taxonomy . . . . . 53
5.6 Conclusions and next steps . . . . . 60

**6 Conclusion** **61**

6.1 Limitations of the current taxonomy . . . . . 61
6.2 Advantages of a new, bottom-up taxonomy . . . . . 61
6.3 Challenges for developing a new taxonomy . . . . . 62
6.4 Next steps . . . . . 62
6.5 Implementation considerations . . . . . 62

# 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 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 some of its 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 a company's SIC code based on its description, and potential explanations for model error that might be suggestiive 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 SIC4s 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 creating networks of words based on their co-occurrence in company descriptions and decomposing them into 'communities' of densely connected terms (communities of industrial language) indicative of 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, or the development and adoption of emerging techbologies. A downside of this approach is that our pipeline is complex and requires tuning at multiple stages, creating potential concerns about robustness and explainability of results. We also find significant challenges(i.e. high error rates) when we try to label companies into the categories of this new taxonomy.

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 a 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 mutually exclusive and completely exhaustive, meaning that 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' (Mateos-Garcia, Stathoulopoulos, and Thomas 2018)

[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.

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, prodesses 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 network analysis techniques to identify "new sectors" as communities of frequently co-occurring terms in company descriptions. We then create a labelled dataset of companies with a strong presence of terms belonging to specific "new sector communities" and then classify other companies into these new sectors based on their semantic similarity to them.

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 industrial taxonomies.

## 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

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

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 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

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

(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: count-vectorising names to produce a matrix for each set of names (rows are names, columns are tokens); {Term frequency}-{inverse document frequency} (TFIDF) transforming our matrices (considering both datasets of names to come from the same corpus for this purpose); 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 excelling at capturing 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,

generate a network of words where the strength of connections between words is based on how frequently they co-occur in documents. 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).

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 (uppercased):

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

- 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.

# 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 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

The model we build is 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

[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.

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 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.

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.

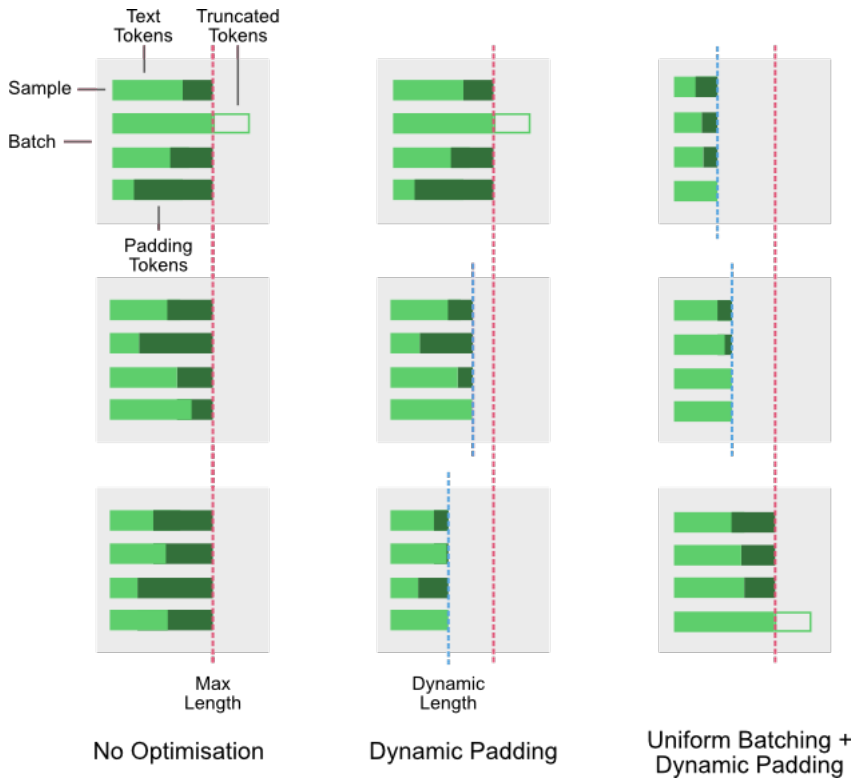


Figure 1: With dynamic padding and uniform batching, texts of similar length are grouped together and padded only to the length of the longest sequence, ensuring that all batches, and therefore all sequences, take a minimal time to process.

## 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 tradeoffs 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.

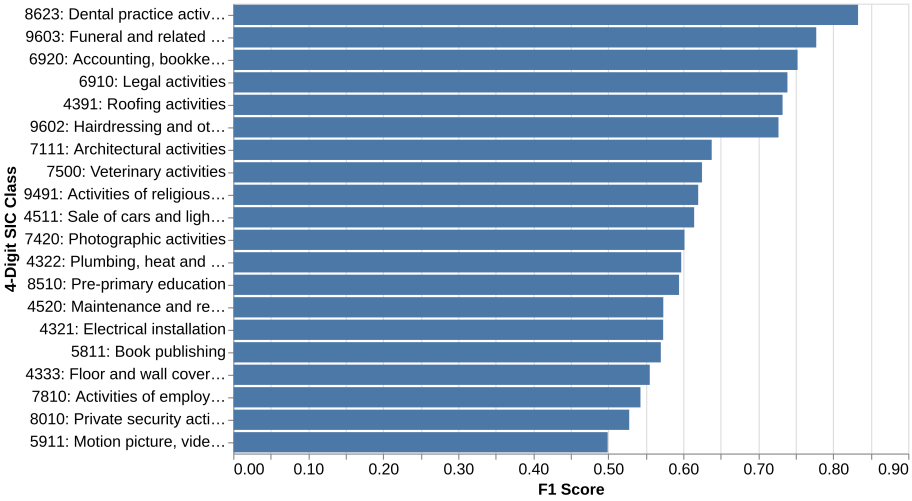


Figure 2: Top 4-digit SIC codes according to their F1 scores.

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 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.

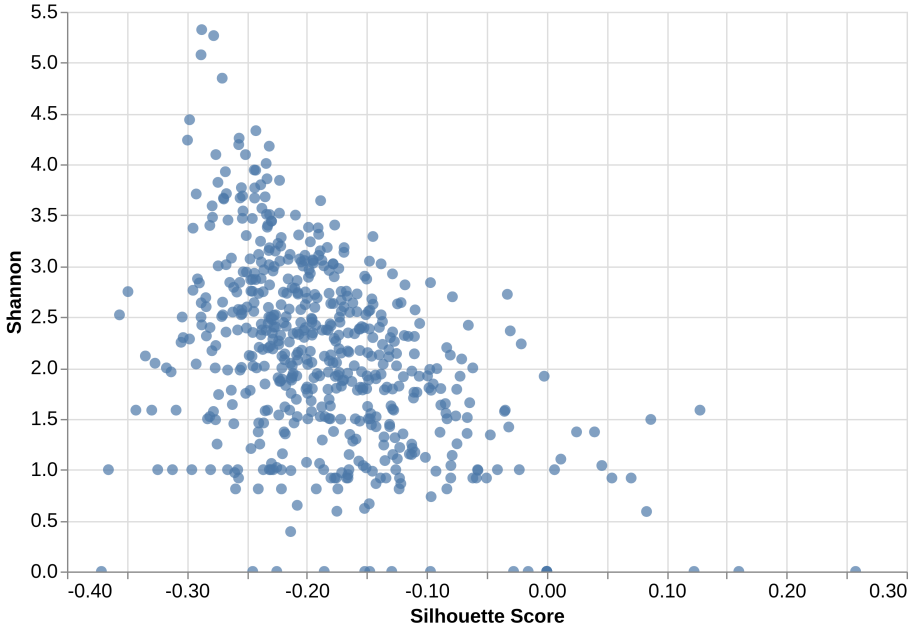


Figure 3: Comparing the silhouette score of company description embeddings and Shannon index of predictions for each 4-digit SIC code shows that SIC codes with less well defined semantic space are more likely generate a diverse array of predictions.

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 orignally 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 |
| 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).

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.

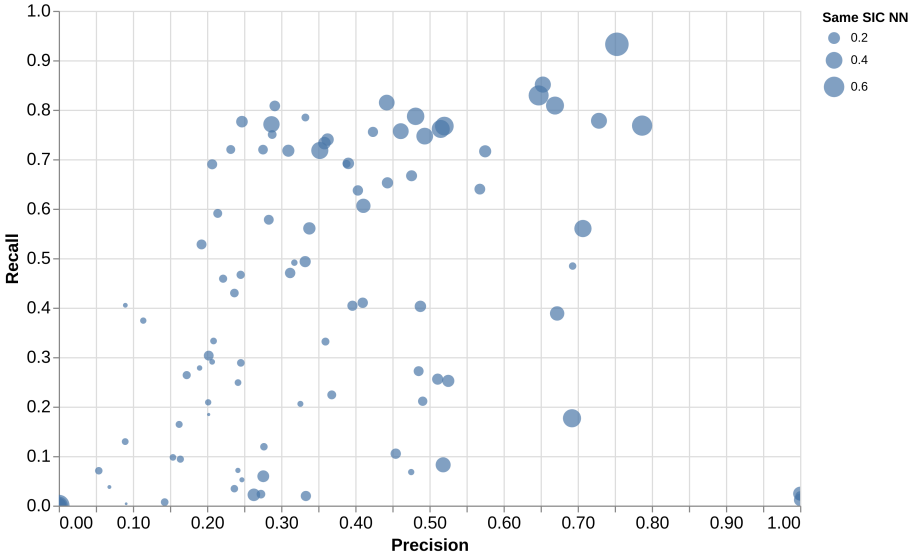


Figure 4: The precision and recall metrics show that some SIC codes are dominated by different classification characteristics, likely to be driven by the degree and type of semantic overlap with companies in other codes.

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). 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.

## 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 by which 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.

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 5, 6, 7 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.

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

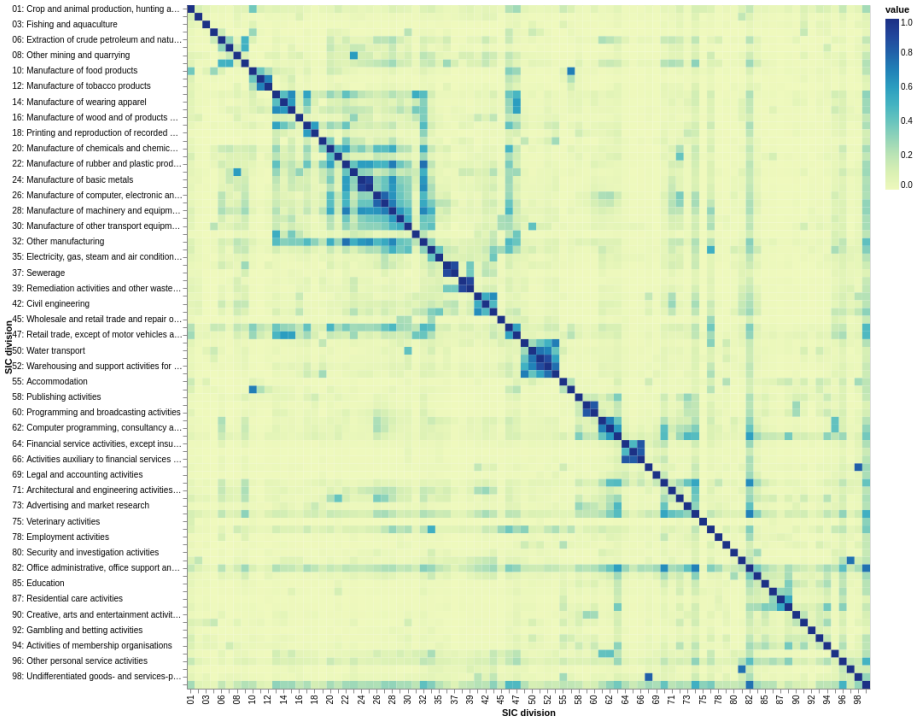


Figure 5: Cosine similarity between SIC divisions similarities (Level 0)

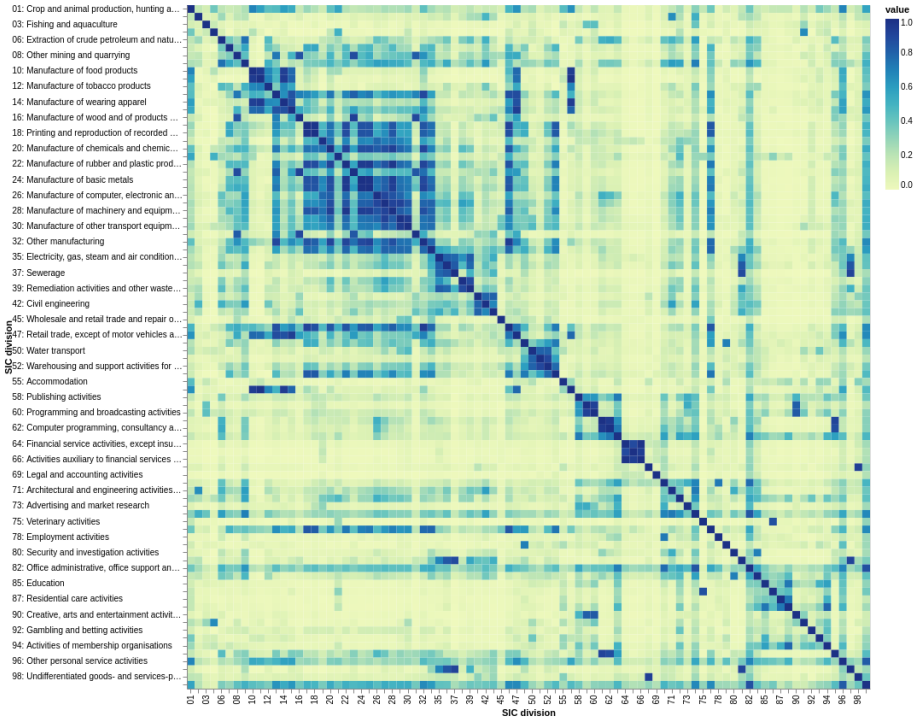


Figure 6: Cosine similarity between SIC divisions similarities (Level 1)

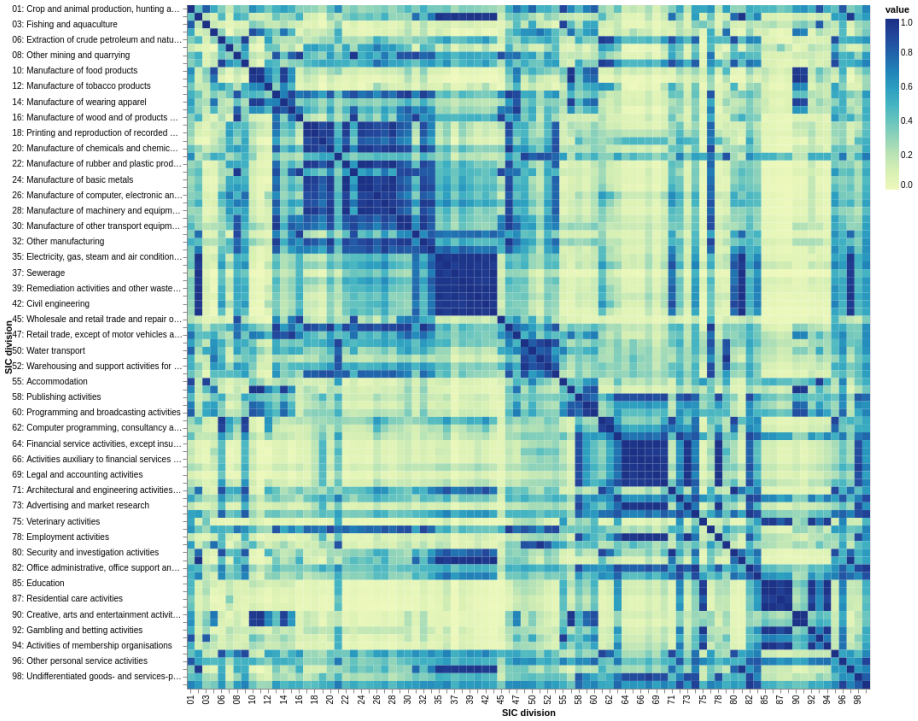


Figure 7: Cosine similarity between SIC divisions similarities (Level 2)

The diagonal block-structure (particularly visible in figure 7) 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 |
| 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 doesn't 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 8 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.

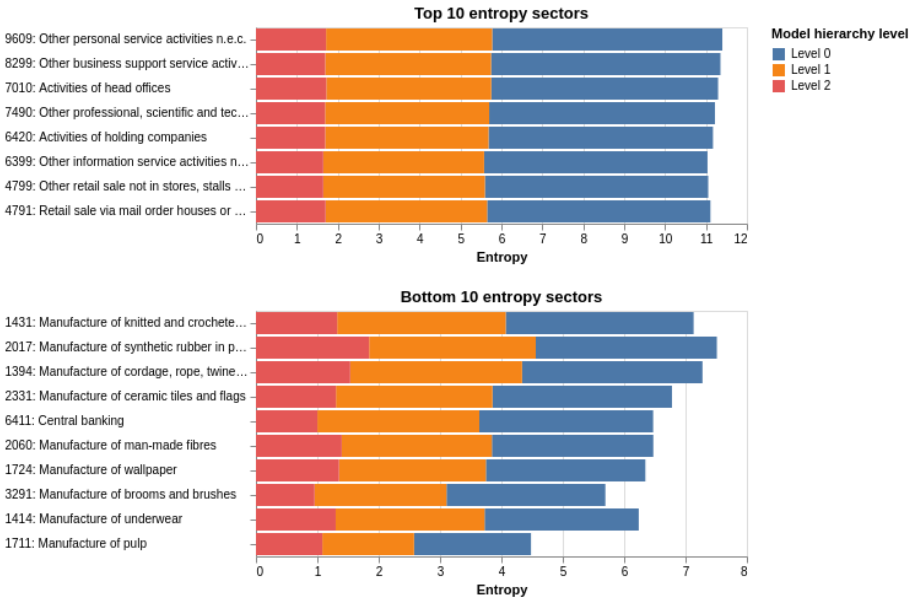


Figure 8: Ten most heterogeneous (highest entropy) and ten least heterogeneous SIC classes calculated according to the entropy of the topic distributions of each SIC class.

Figure 9 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.

## 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

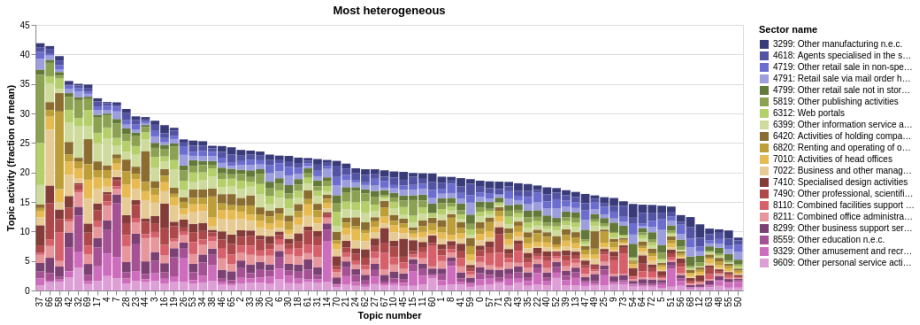


Figure 9: Distribution of topic activity amongst the top 20 most heterogeneous (according to entropy of the topic distributions with SIC codes) sectors. Activity for a topic-sector combination is expressed as a fraction of the mean activity of that topic across all sectors. Based on level 1 of the model hierarchy.

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.

# 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 the mismatch between business descriptions and codes for some of the SIC codes, the strong overlap between codes in different parts of the taxonomy, and the presence of semantically heterogeneous 'not elsewhere classified' codes that attract very different types of economic activity as well as some companies that could be classified elsewhere in 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]

[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.

2. Its production is automated: The categories within the taxonomy are identified and named automatically with some manual tuning of key parametres to increase the intepretability and relevance of outputs.
3. It can be used to tag companies with laabels 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 SIC4 codes, and, potentially to 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 have to build a 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 10.

## 5.1 Methodological narrative

We think of industrial sectors as a set of latent constructs $I = \{i_1, ..., i_n\}$ where industry $i_k$ is manifested in a vocabulary of words $S_k = \{w_1, ..., w_m\}$ about products and services, production processes and technologies that happen with a high probability in the description of companies that operate in it. As an example, the video games sector may be comprised of a vocabulary including *console*, *game*, *video game*, *adventure*, *role playing game*, *app* etc.

I addition to this set of sector specific words $S$, there is a set $G$ of generic words that are uninformative about the sector that a company operates on. This may include general business terms as well as marketing language used to promote products and services online.

The challenge that we face is to infer $I$ from the distribution of $S$ and $G$ in our corpus of company descriptions. In order to do this, we would like to (1) remove from company descriptions all $w_i \epsilon G$ that are uninformative about the sector where a company operates, and to (2) identify clusters of words such that their words $w_i \epsilon S_k$. The text processing stage of our analysis sets out to achieve (1) and the community extraction stage sets out to achieve (2).

More specifically, in (2) we construct a network of words based on their co-occurrences in companies and deconstruct that network into a set of communities $C = \{c_1, ...c_n\}$ composed of densely connected (frequently co-occurring) words. We assume that $C_k \sim S_k$ - each of the resulting communities of words is a representation or proxy of latent sector $k$. As noted in the first feature of our taxonomy, we undertake this community detection *inside* each SIC4 code.[9]

[9] Undertaking our community detection at the SIC-4 level also gives us a ready-made way to remove generic business terms - those that appear in many different SIC4s - and makes the community detection effort more computationally tractable by reducing the size of the networks we work with.

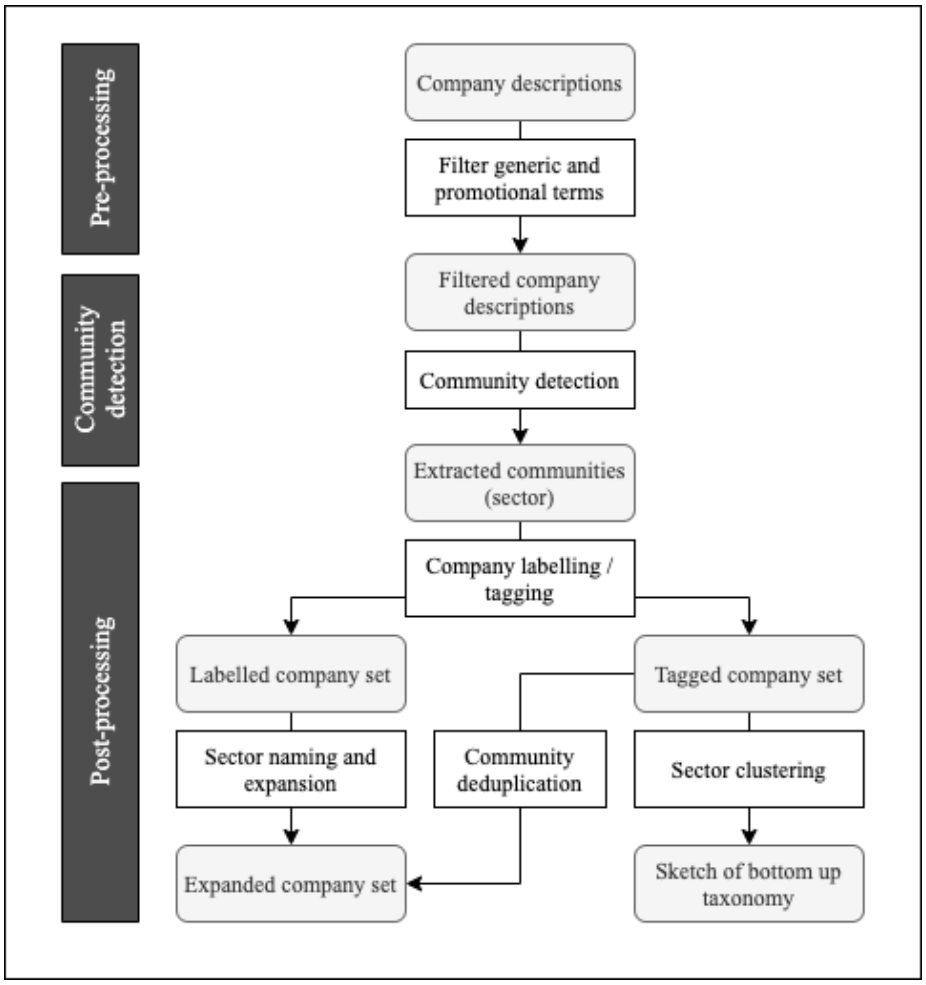


Figure 10: Summary of our pre-processing, analysis and post-processing pipeline

Having identified a set of word communities that proxy 'sectors,' we move into a company labelling stage where we identify a set of companies $F_k = \{f_1, ..., f_n\}$ such that a company $f_j \epsilon F_k$ if $w(f_j) \cap S_k > \tau$, where $w(f_j)$ is the set of words in the description for company $f_j$, and $\tau$ is a critical value. We also use the salient terms in each of these labelled sets to name the community $C_k$ and calculate the semantic homogeneity of those terms to proxy a community's coherence (that is, we consider communities whose salient terms are more similar to each other to be more coherent than communities composed of heterogeneous salient terms).

In the post-processing stage, we clean and expand our list of sectors and labelled companies.

The labelling approach described above makes it possible for us to tag a company with multiple sectors that occur frequently in it. Some of these co-occurrences may reflect situations where a company straddles sectors (e.g. a fintech company tagged with a financial sector label and a software sector label) but in others it may simply capture sector duplicates across SIC4 (i.e. situations where due to misclassification a subset of a sector $I_k$ is placed in SIC4 $A$ and another subset in SIC4 $B$). We merge duplicated communities and replace them in our tagged company set.

One limitation of our labellling strategy is that it will only capture companies that include the exact words in a sector vocabulary $S_k$, potentially excluding relevant companies that use synonyms or are below our hard threshold $\tau$. In order to improve the recall of our approach we vectorise all company descriptions in our corpus and calculate the distance between the vector representation of each unlabelled company and a vector that summarises each of our sectors $i_k$ with the average of the vectors for all companies classified in that sector, $F_k$. We assign each unlabelled company $u$ to $F_k \mid d_{l,i} = min(d_{l,i \epsilon K})$ if $d_{l,i} < \rho$, another critical value.

The rest of the section goes through these steps providing additional detail about our implementation of relevant methods and emerging results. Before doing this we note that in the text we will refer interchangeable to communities (of words) that represent sectors, and the sectors that compose the bottom-up taxonomy we are developing.

## 5.2 Text pre-processing

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`).[10]

We now want to subject this corpus of company descriptions to additional processing in order to generate individual company descriptions primarily comprising sector-specific ($S_i$) terms. To achieve this, we:

### 5.2.1 Filter low sector salience terms

We concatenate all tokenised company descriptions inside each SIC4, calculate term frequency per SIC4 and calculate their TF-IDF score (this roughly normalises the occurrence of a term in the SIC4 code over its occurrence in the whole corpus) and keep, for each SIC4, only those terms in the top 5% TFIDF score.

We find that even after taking this step, the resulting SIC4 vocabularies still contain a large number of generic business terms - this is because word frequencies are highly skewed, with a small number of frequently occurring terms that overlap between sectors. To adress this, we remove all terms that are present in the salient vocabularies of more than 50% of SIC4s.

Table 6 compares a random set of 10 terms in selected SIC4s before and after the application of this additional filter - we note that a significant number of non-sector specific terms remain. We have decided against setting a more strict threshold for exclusion of salient terms because these resulted in a significant decrease in the size of company-level vocabularies, particularly for "not elsewhere classified" SIC codes whose salient terms are more evenly distributed across the corpus of business descriptions.

Table 6: Randomly sampled terms extracted from sector vocabularies after applying filter 1 (only keep terms with tfidf score in top 5% for sector) and filter 2 (removal of terms that are present in more than 50% of sectors)

| SIC4 | Filter 1 | Filter 2 |
|---|---|---|
| 0123 | green, long_term, having, population, standard, unique, business, local, processing, strawberry, add_value, salad, security, produce, bakery, custom, people, ingredient, farm, successful | garden, fruit_vegetable, farm, cut_flower, able_provide, simon, add_value, fruit, paradise, hope, future_generation, custom, employ, bit, bakery, live, long_term, formulation, ingredient, having |
| 6201 | culture, turn, popular, late, place, multi, aim_provide, help, small_business, seek, dedicate, security, demand, perform, help_achieve, tech, know, crm, standard, mind | reflect, journey, low, forefront, visual, critical, goal, optimise, stay, understanding, speak, generation, suit, explore, utilise, intuitive, form_date, creative, legal, expect |

[10] In the text that follows we will refer to these tokens, which may be unigrams, bigrams or trigrams, as 'terms.'

| SIC4 | Filter 1 | Filter 2 |
|---|---|---|
| 7490 | page, private, loc_loc, original, total, identify, line, exist, qualification, innovative, assess, sale, relate, base, experienced, policy, operate, able_provide, global, offer_range | explore, etc, website, prepare, example, demonstrate, appointment, selection, learn, plant, plan, detailed, necessary, healthcare, good_practice, consultancy, talent, maximise, specification, scheme |
| 9999 | financial, kind, enable, rely, run, accredit, consultant, campaign, study, transport, whilst, idea, industrial, quote, class, enjoy, time, found_date, locate_gpe, cut | think, website, sign, packaging, offer_range, turn, gift, outside, flexible, early, specification, member_org, practice, student, travel, original, save, successfully, clear, direct |

### 5.2.2 Removing marketing tems

We apply an additional filter to remove promotional and marketing terms that companies use to advertise their products and services in their websites but are uninformative about the sector where a company operates.

In order to do this, we use VADER (Valence Aware Dictionary for Sentiment Reasoning), a sentiment analysis model which scores words based on their polarity (whether they express a positive or negative sentiment) and intensity (Hutto and Gilbert 2014). We apply the NLTK implementation of VADER to every term in our corpus and remove those with a high polarity score (above 0.5) (Loper and Bird 2002). This leads us to remove 720 terms from our vocabularly. figure 11 displays the highest vocabulary terms in our vocabulary.

In figure 12 we present the mean number of promotional terms per SIC-2 (division). We have removed the labels from the bottom to increase readability and added the codes for the top five divisions in the table - they are consumer-oriented products and services such as Accommodation (55), Food and beverage service activities (56), Travel agency, tour operator and other reservation service and related activities (79), Education (85), Manufacture of food products (10).

Figure 13 summarises the results of our filtering exercise: we reduce the mean number of terms in company descriptions significantly (by 72% on average). This is particularly visible for Not Elsewhere Classified sectors (cross shapes). The chart also illustrates the stronger presence of services sectors in our dataset.

## 5.3 Community extraction

Having filtered the terms in company descriptions, we move to the creation of a term co-occurrence network inside each SIC4, and its decomposition into communities representing sectors, focusing on 80 SIC4s in our dataset that contain more than 500 companies.

Figure 11: Wordcloud of the top 100 positive terms in business descriptions

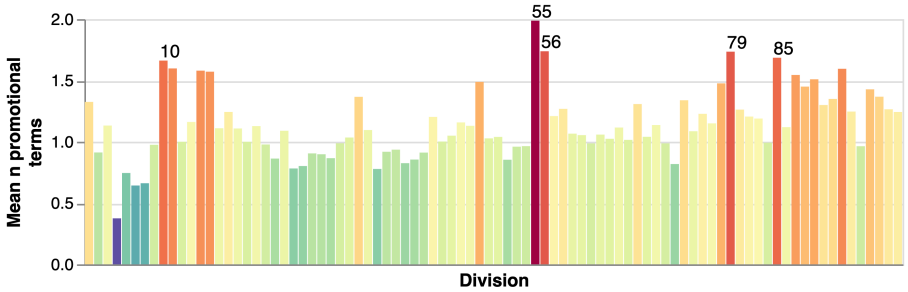


Figure 12: Mean frequency of promotional terms by SIC divisions, with top five codes highlighted

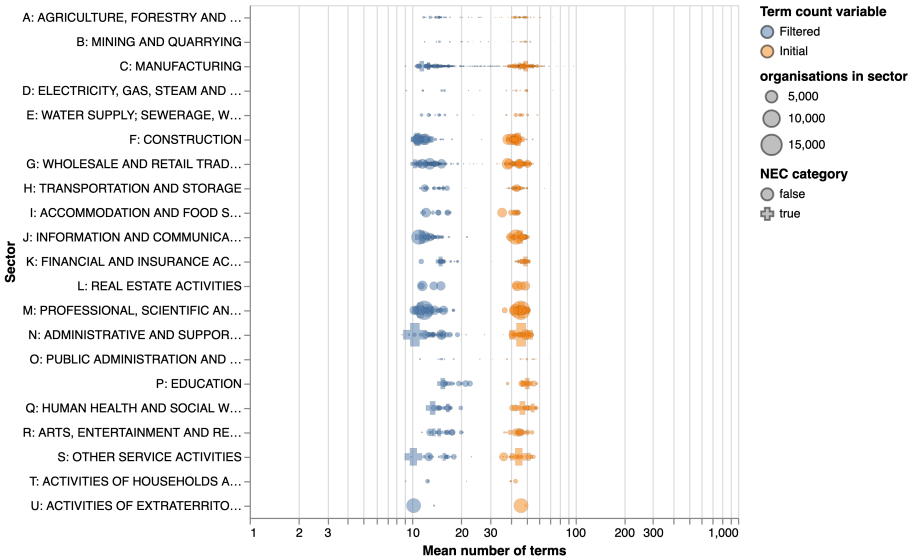


Figure 13: Mean term count for SIC4 industries in different SIC divisions before and after filtering. The size of the point indicates the number of companies in a sector, and the shape whether it is a "Not elsewhere classified" sector (cross) or not (circle)

### 5.3.1 Creation of co-occurrence networks

Each term co-occurrence network comprises a collection of nodes $N = \{n_1, ...n_m\}$ where $n_i$ is a term within our filtered set, and a collection of edges $E = \{e_1, ...e_k\}$ with weights $W = \{w_1, ...w_k\}$. The value of the weights represent the number of instances that two nodes $n_i$ and $n_j$ co-occur in a company description in the SIC4.

We treat terms that occur in more than 20% of the company descriptions as 'stop terms' which are less useful for our community detection purposes, and simplify the network. This involves removing edges with low weights by extracting their maximum spanning tree - the set of nodes and edges in $N$ that preserves a fully connected network with the maximum sum of weights (Asano et al. 1988). This has the goal of making the network sparser while preserving the strongest connections that may be indicative of community structure. We add to this network the top 20 edges by weight from the original network that had been removed by the maximum spanning tree.[11]

[11] This has the goal of preserving more structure from the original network. We preserve a relatively low number of edges in order to reduce the computational complexity that we have to tackle in the community detection stage of analysis.

### 5.3.2 Community detection

Community detection (or discovery) techniques seek to learn the structure of networks by partitioning them into groups of nodes that are densely connected internally and less connected with nodes in other communities (Fortunato and Hric 2016, javed2018community).

There are dozens of community detection algorithms developed in many different disciplines and we do not have strong priors that a particular algorithm will be specially effective for decomposing our term co-occurrence networks so we adopt a grid-search strategy to tackle our challenge. This entails selecting an ensemble of algorithms and range of potential parameters, using them to partition our networks, and comparing the modularity of the results. [tbl:4_2] summarises the ensemble of community detection algorithms that we have used, as implemented in the `cdlib` python library (Rossetti, Milli, and Cazabet 2019).

Table 7: Community detection algorithms implemented in our analysis

| Algorithm | Reference |
|---|---|
| Louvain | (Blondel et al. 2008) |
| Leiden | (Traag, Waltman, and Eck 2019) |
| AGDL | (Zhang et al. 2012) |
| Chinese whispers | (Ustalov et al. 2019) |
| Label propagation | (Raghavan, Albert, and Kumara 2007) |
| Markov clustering | (Enright, Van Dongen, and Ouzounis 2002) |

Figure 14 presents the modularity scores achieved by different community detection algorithms by SIC4, and the size of the communities extracted.[12] In general, the Leiden and Louvain algorithms, which recursively assign nodes to communites with the goal of optimising the modularity score perform most effectively while extracting a relatively small number of communities. Perhaps as expected, we find that "not elsewhere classified" SIC4s tend to achieve more modular network partitions, suggesting a higher degree of heterogeneity and scope for decomposition into additional subsectors.

### 5.3.3 Robustness

Do the best performing community detection algorithms tend to partition terms in the same communities? To evaluate if this the case we calculate, for each SIC code the normalised mutual information between the partitions generated by pairs of algorithms in the top three by modularity score.[13]

[12] We use the Erdos Renyi modularity score, which captures the extent to which intra-community connections happen at a higher rate than would be expected if nodes were connecting randomly (McDiarmid and Skerman 2018)

[13] Normalised mutual information captures the correlation between two partitions, with a score of 1 for perfect correlation and a score of 0 for a independence.

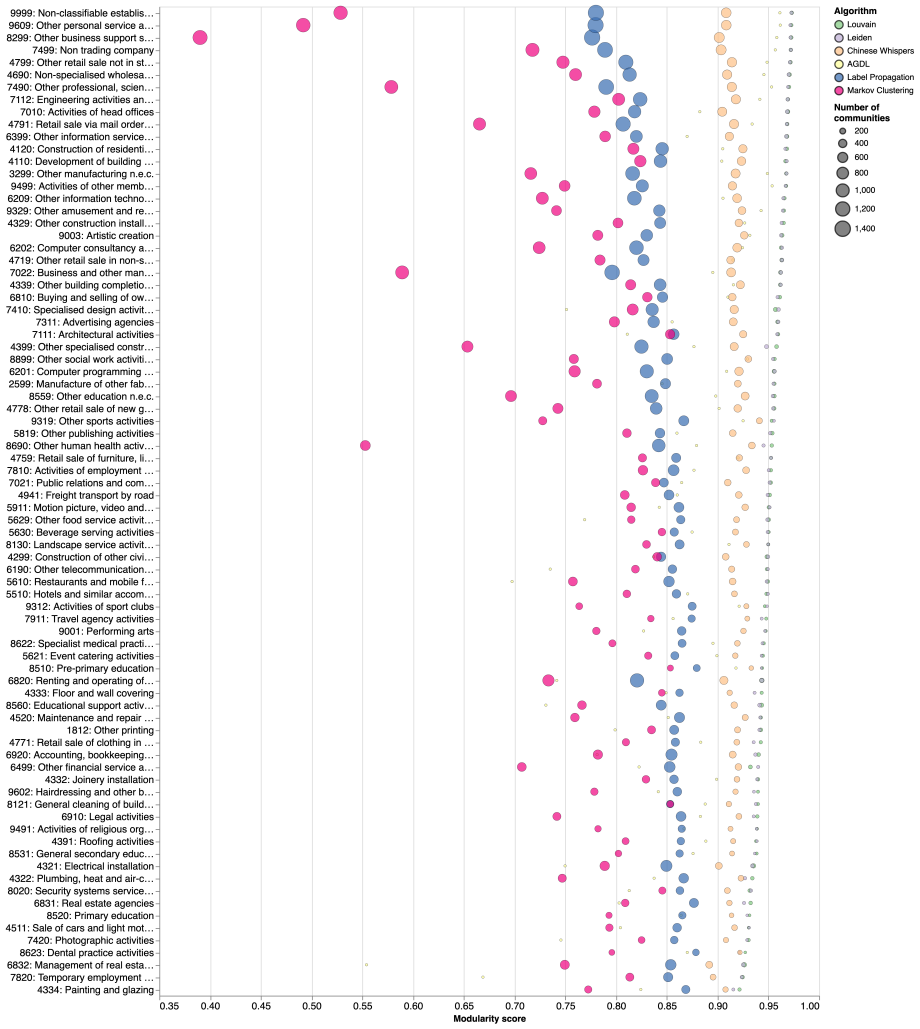


Figure 14: Modularity scores for SIC4 community partitions by community detection algorithm. The size of the points indicates the number of communities extracted. We sort the vertical axis by the maximum modularity achieved in the partition

As figure 15 shows, the normalised mutual information scores tend to be high (between 0.75 and 0.98), suggesting that the outputs of our community detection efforts are robust to the choice of algorithm. We note that sectors in "Other" / "Not elsewhere Classified" sectors in services and retail tend to have lower median scores, suggesting that it is harder to classify terms into communities for those sectors.

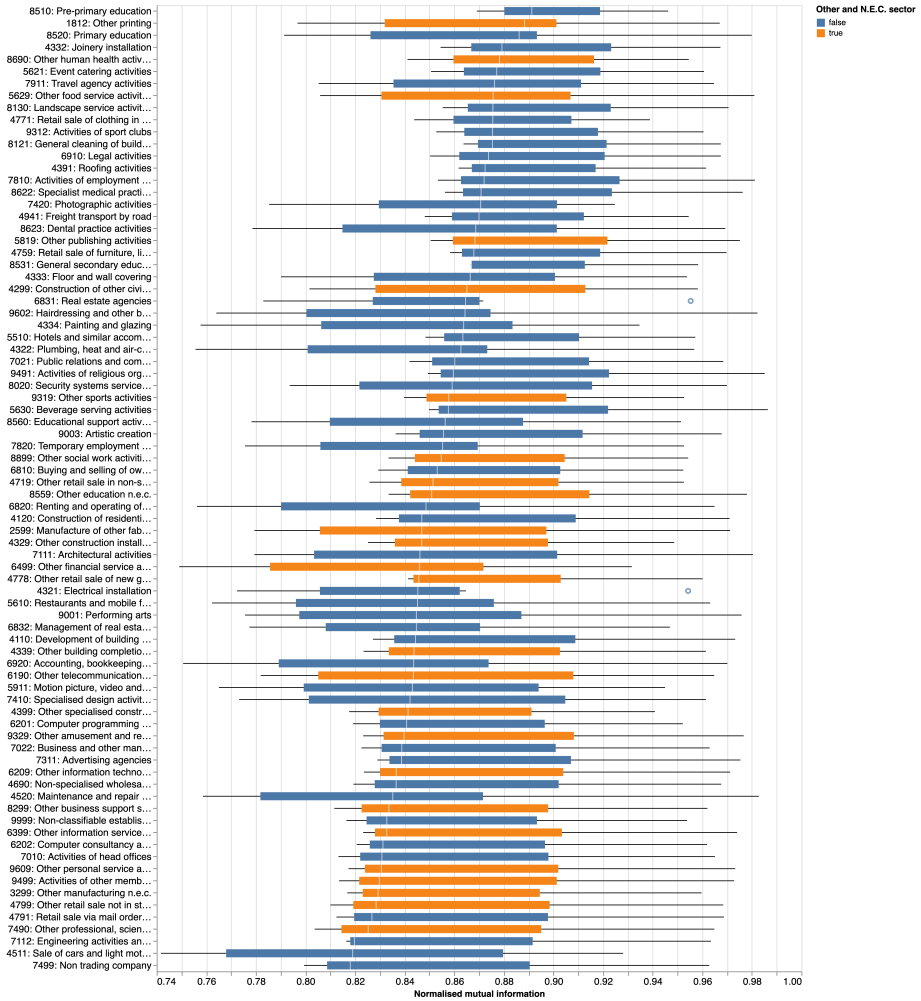


Figure 15: Distribution of partition similarities based on normalised mutual information for top performing algorithms by SIC4. The colour of the bar indicates if a SIC4 is in an "Other" and/or "Not Elsewhere Classified" sector

## 5.4 Post-processing

Our community detection effort yields outputs that require additional processing in order to remove noise, increase interpretability and eventually classify companies into sectors. As an illustration, we have extracted 6247 communities from 80 SIC4. As figure 16 shows, many of these communities include very few terms, and several of them include many terms. We would optimally like to remove communities that capture few firms and duplicated communities, to assess their semantic coherence (the extent to which they include terms referring to similar activities) and remove incoherent ones, and to name in a way that makes them more interpetable.

The approach we take is to use the terms assigned to different communities in order to label companies into sets that define sectors. We then analyse these labelled sets to extract sector names, identify community duplicates and related companies. The rest of the section outlines our methods and results.

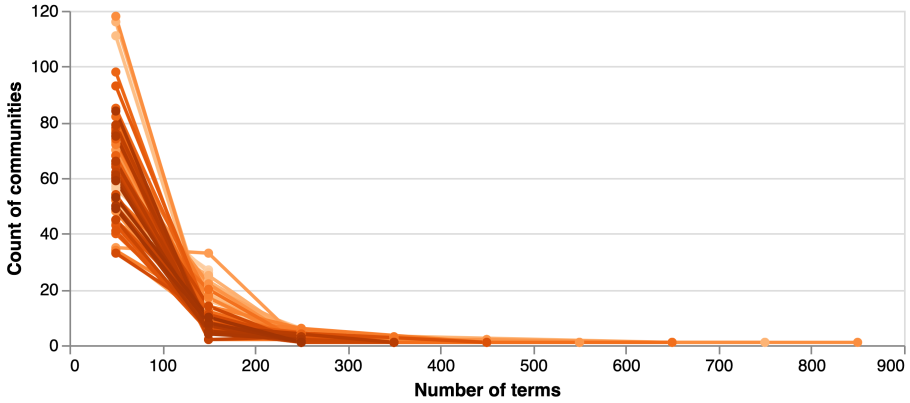


Figure 16: Histogram of term counts inside each community in a SIC4: Each line represents a SIC4, with the frequency of various term counts in the communities extracted for the SIC4

### 5.4.1 Company labelling

We begin by counting the overlap between the set of terms in a company description and a community. We opt for doing this instead of counting the raw number of occurrences of community terms in company descriptions in order to reduce the impact of common and generic terms. Here, our analysis branches into two paths:

- We assign each company to the community with the biggest overlap between the set of terms in its description and the community vocabulary but only if this overlap consists of five terms or more.[14] This yields a labelled dataset

[14] This means that any community consisting of less than five terms is, by definition, excluded

with 50,700 companies in 5117 unique communities that we use to name sectors. We will later expand this labelled dataset with companies that are semantically similar to the companies that we have classified in various sectors.

- We will use the tagged dataset where each company has a tag for each sector where company description and sector vocabulary overlap above the critical value *tau* (4 occurrences) to deduplicate communities and build the bottom-up taxonomy later in the analysis.

#### 5.4.2 Community naming and coherence

Similar to what we did at the text pre-processing stage, we identify salient terms in each of our groups of labelled communities by concatenating company descriptions *inside those labelled sets* and identifying the ten terms within each of the sets with the highest TF-IDF scores - we think of them as the community (sector) names. As an additional step, we calculate these names' coherence by calculating the median pairwise distance between a vector representation of each of their constituent terms based on a `word2vec` model trained on the corpus (Mikolov, Yih, and Zweig 2013). Sectors with high coherence tend to have similar terms in their names, and sectors with low coherence tend to have different tems, which could indicate low-quality communities.

Table 8 presents some examples of high coherence, medium coherence and low coherence communities. The examples show that high coherence communities capture highly specific sectors such as dental health (`8263_6`), golfing (`9312_0`) or wedding dresses (`4771_1`). We note that less coherent communities are not always low quality. For example, community (`9999_38`) seems to capture musical performances, and community (`5629_5`) wine-related activities. Having said this, later in our analysis we will focus on relatively coherent communities with a coherence score above 0.5.

Table 8: Examples of communities with different levels of name coherence

| n | Community | Salient terms | Median Similarity | Ranking |
|---|---|---|---|---|
| 0 | 8623_6 | dental_health, dental, tooth, dentist, dentistry, visit_dentist, hygienist,... | 0.889 | high |
| 1 | 9312_0 | cardinal_hole, golf_course, golf_club, golf, par_cardinal, lounge, clubhous... | 0.851 | high |
| 2 | 4771_1 | wedding_dress, bride, dress, bridal, gown, bridal_gown, wedding_day, seamst... | 0.835 | high |

from the rest of our analysis.

| n | Community | Salient terms | Median Similarity | Ranking |
|---|---|---|---|---|
| 3 | 7499_38 | meat, sausage, butcher, butchery, alliance, pork, table_tennis, net, lamb, . . . | 0.835 | high |
| 4 | 9609_58 | payroll, accountant, accounting, accountancy_service, sole_trader, tax, tax. . . | 0.827 | high |
| 5 | 9999_46 | solicitor, family_law, legal_advice, solicitors, dispute, employment_law, p. . . | 0.824 | high |
| 6 | 9999_29 | tax, tax_return, payroll, sole_trader, self_assessment, bookkeeping, tax_pl. . . | 0.820 | high |
| 7 | 8622_12 | chiropractor, chiropractic, pilates, clinic, low_pain, pain, motion, sympto. . . | 0.788 | high |
| 8 | 6910_4 | immigration, immigration_law, immigration_advice, visa, visa_application, l. . . | 0.774 | high |
| 9 | 9609_27 | tenant, landlord, letting, landlord_tenant, residential_property, tenancy, . . . | 0.757 | high |
| 10 | 7911_22 | cottage, self_cater, sightseeing, hostel, picturesque_village, self_caterin. . . | 0.714 | medium |
| 11 | 8130_1 | tree_surgery, tree_surgeon, hedge, arboriculture, aspect_tree, landscaping,. . . | 0.713 | medium |
| 12 | 1812_23 | carrier_bag, quick_turnaround, dispose_waste, existing_new, builder_archite. . . | 0.711 | medium |
| 13 | 9602_5 | therapist, therapy, treatment_room, mind_body, leave_feel, pamper, salon, s. . . | 0.671 | medium |
| 14 | 9312_4 | coaching, cricket, camp, tennis, expedition, gymnast, polo, participant, sq. . . | 0.625 | medium |
| 15 | 8559_6 | surgical, surgery, surgeon, acupuncture, nurse, bursary, clinical, medicine. . . | 0.621 | medium |
| 16 | 4339_15 | suspend_ceiling, ceiling, partition, fire_protection, partitioning, corrosi. . . | 0.588 | medium |

| n | Community | Salient terms | Median Similarity | Ranking |
|---|---|---|---|---|
| 17 | 5819_3 | media, radio, photography, video_content, footage, advertising_agency, vide... | 0.577 | medium |
| 18 | 4322_12 | drain, drainage, pipe, plumbing, guttering, block_drain, high_pressure, rod... | 0.539 | medium |
| 19 | 2599_6 | laser, fold, laser_cut, blank, quantity_length, thickness, invest_late, bro... | 0.501 | medium |
| 20 | 6209_63 | investigation, criminal, testimony, forum, virtual_office, repository, indi... | 0.444 | low |
| 21 | 7410_57 | collective_knowledge, sound_understanding, think_tank, legal_profession, to... | 0.410 | low |
| 22 | 9999_38 | sing, opera, singing, breakthrough, consortium, performer, leadership_devel... | 0.404 | low |
| 23 | 5819_52 | bookshop, vale, follow_acquisition, wo_nt, direct_mail, invoicing, strategi... | 0.324 | low |
| 24 | 8690_74 | computing, academic, entrant, postgraduate, bsc, msc, hold_relevant, curren... | 0.323 | low |
| 25 | 4759_25 | vintage, clock, outdoor_living, wide_array, dimension, mechanism, transform... | 0.300 | low |
| 26 | 5629_2 | wine, dinner, wines, alcohol, vineyard, wine_list, final_product, necessari... | 0.273 | low |
| 27 | 7112_20 | pump, oem, bike, acclaim, casting, transmission, switchgear, fundamental_pr... | 0.217 | low |
| 28 | 8020_54 | weatherproof, phenomenon, damp, frequent, decorate, painting, portrait, por... | 0.217 | low |
| 29 | 5629_64 | predictable, linen, steady_growth, textile, quality_material, aat, portfoli... | 0.182 | low |

Figure 17 displays the distribution of community coherences by SIC code sorted vertically by the share of all communities extracted from a SIC4 that have a

high degree of coherence (above 0.7). The only pattern worth noting is that seeemingly less knowledge intensive SIC4s tend to have a larger share of high coherence communities, suggesting a higher degree of specialisation within their constituent sectors. We see many SIC4s related to creative and information activities at the bottom of the Y axis, suggesting less coherent, more diversified sectors.

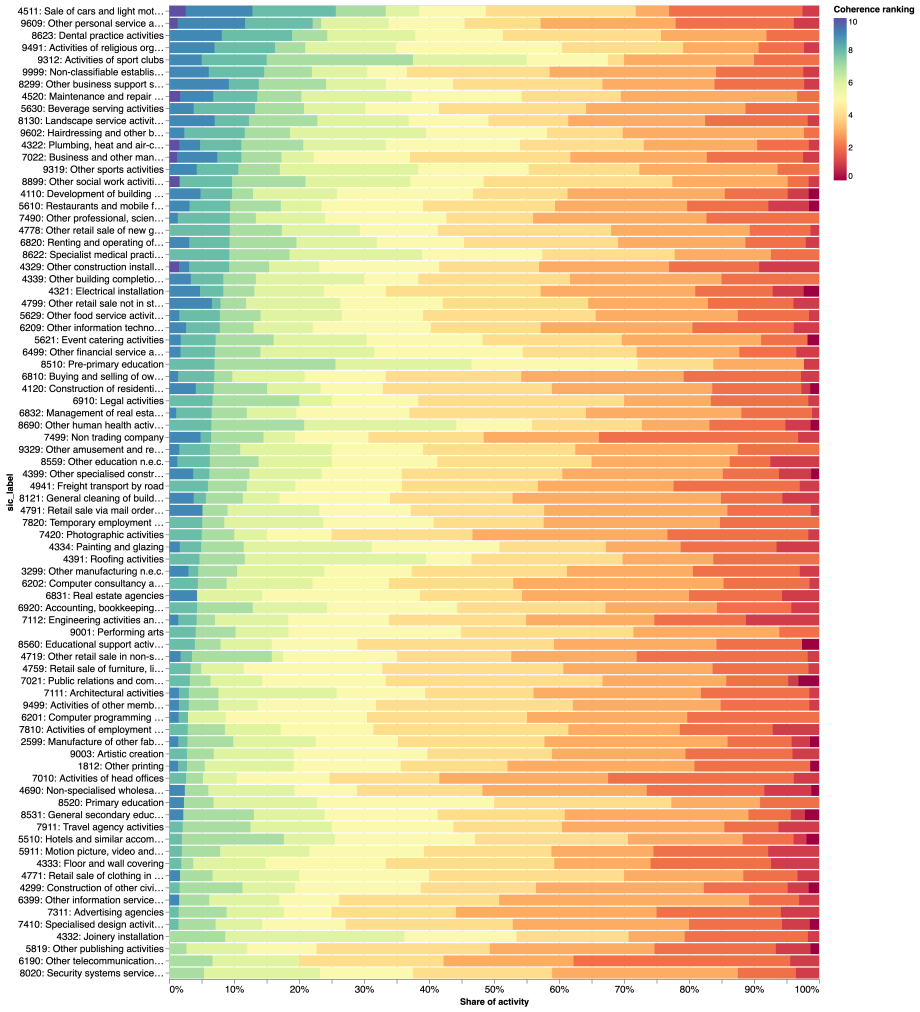


Figure 17: Coherence distribution by SIC4: Blue tones indicate the share of more coherent communities, and orange tones the share of less coherent communities.

### 5.4.3 Examples

Figure 18 and figure 19 display illustrative, simplified term networks and their partitions based on 1000 randomly sampled companies in four sectors (8299: Other Personal Services Not Elsewhere Classified 7490: Other professional, scientific and technical activities Not Elsewhere Classified, 6201: Computer Programming Activities and 4334: Painting and glazing). The colour of the nodes represent the communities that various terms belong to, focusing on the top twenty communities inside each SIC4.

We note that the communities for 8299 and 7490, two highly heterogeneous 'not elsewhere classified' industries) include a wide range of activities including, in the case of 8299, services ranging from health to education and entertainment to energy. The term network for 6201 appears to capture different segments of the software industry (web design, apps, financial services, accounting and customer-relationship management systems) as well as emerging technologies such as machine learning and virtual reality. The communities in 4334 are much less distinct, suggesting more homogeneity and less scope for partitioning that sector into a more granular set of industries.

Table 9 provides some examples of communities including terms that relate to sustainability, renewable energies and the environment. It shows that our method generates categories that might be useful for the analysis of 'green industries' that are not currently included in the SIC taxonomy, and to study the adoption of green and renewable methods in other sectors such as construction, waste management or finance.

Table 9: Example communities mentioning terms related to the environment, renewables and sustainability

| community name | salient_terms |
|---|---|
| 2599_51 | flue, coupling, systems, storage_distribution, ce_marking, testing_inspection, environmentally_aware, production_distribution, hose, operate_independe.. |
| 4120_1 | renovation, list_building, house_extension, loft_conversion, brick, period_property, repeat_customer, utmost_importance, environmental_impact, buildin.. |
| 4321_20 | heat_pump, boiler, air_source, central_heating, gas_oil, plumbing_heating, ground_source, hot_water, shower, renewable.. |
| 4321_30 | reduce_energy, consumption, interior_design, environmental_performance, environmental_policy, environmental_impact, minimise_waste, environmental_legi.. |
| 4321_5 | renewable_energy, turbine, renewable, wind_turbine, green_energy, biomass_boiler, project_manager, electrician, renewable_technology, boiler.. |

| community name | salient_terms |
|---|---|
| 4321_82 | instrumentation, pharmaceutical, monitoring_system, equate, cooling, data_centre, dramatically_reduce, variant, food_processing, energy_consumption.. |
| 4322_11 | renewable_energy, solar, greenhouse_gas, solar_energy, heat_pump, solar_thermal, certification_scheme, carbon_dioxide, plumbing_heating, climate_chang.. |
| 4322_22 | heat_recovery, heat_exchanger, waste_water, difficult_time, energy_saving, usage, funeral_plan, ductwork, plasma, mechanical_engineer.. |
| 4322_6 | heating_system, boiler, biomass_boiler, warm_air, biomass, wood_chip, plumbing, ecological, renewable, episode.. |
| 4322_9 | heat_pump, ground_source, air_source, renewable_heating, renewable_technology, renewable_energy, renewable, renewable_heat, boiler, solar_panel.. |
| 4329_0 | insulation, cavity_wall, food_waste, thermal_insulation, energy_saving, industrial_commercial, green_deal, bill, water_softener, energy_bill.. |
| 4329_11 | heat_pump, renewable_energy, solar_pv, solar_thermal, air_source, renewable, renewable_technology, ground_source, biomass_boiler, renewable_heating.. |
| 4332_12 | double_glazing, heat_loss, double_glazed, energy_bill, heating_bill, extrusion, installers, thermal, window_door, ex_stock.. |
| 4332_18 | carpentry_joinery, environmental_performance, general_building, project_undertake, factory_unit, operate_primarily, responsibly_source, enormous_pride.. |
| 4391_3 | roofing, roofer, roofing_repair, roofing_material, repeat_customer, solar_energy, tile, insurance_back, high_percentage, slate.. |
| 4799_40 | solar, renewable_energy, renewable, low_carbon, energy_saving, solar_panel, solar_pv, solar_energy, wind_turbine, climate_change.. |
| 4941_52 | recycling_waste, wheelie_bin, landfill, bin, environmentally_sound, waste_collection, landfill_site, government_legislation, food_waste, unaware.. |
| 6499_4 | real_estate, private_equity, financing, asset_management, equity, renewable_energy, offset, credit_card, transaction, coal.. |
| 7022_31 | renewable_energy, renewable, energy_saving, reduce_energy, solar, biomass, heat_pump, electricity, fossil_fuel, led.. |
| 7112_4 | renewable_energy, low_carbon, low_energy, carbon, solar, yacht, grid, food_waste, green_energy, problem_solve.. |
| 7410_32 | renewable, pin, loudspeaker, gearbox, wind_turbine, tidal, onshore_offshore, turbine, heat_pump, generator.. |

| community name | salient_terms |
|---|---|
| 7490_29 | green_deal, emission, energy_saving, carbon_footprint, flue, greenhouse_gas, reduce_carbon, refinery, energy_usage, climate_change.. |
| 7490_32 | waste_management, waste_collection, environmental_consultancy, environmental_impact, reduce_waste, waste_recycling, transmission, kick, wheelie_bin, r.. |
| 8130_42 | minimise_waste, environmental_performance, printer_cartridge, electricity_gas, photocopier, environmentally_aware, outgoing, energy_conservation, recy.. |
| 8299_40 | air_conditioning, swimming_pool, heat_pump, design_installation, ventilation_system, spa, reduce_energy, system_instal, installation_maintenance, hot_.. |

#### 5.4.4 Community de-duplication

Our local (SIC4-level) approach to extracting communities creates the risk of duplication. This could happen if for example companies in a sector tend to be classified in several SIC4 codes. In that case, our analysis might extract the same communities of terms in different SIC4s.

We address this by analysing the tagged company dataset mentioned above. Here, each company is labelled with all the communities where it has an overlap in terms above a critical value of 4. We would expect duplicate communities to co-occur in a company. We identify them by calculating pairwise cosine distances between a binary vector of community occurrences in companies and identify the most similar communities (those with a similarity score above 0.7) - we merge these communities and rename them with a combination of the salient terms of their consitutent elements.

Table 9 presents some examples of duplicates: while there are several instances of clear duplication (e.g. between communities capturing financial services and communities related to automobile selling extracted from different SIC4s) in other cases the overlap seems to be linked to non-salient terms within a community vocabulary: for example, a community related to vehicle leasing is merged with a community related to property rental, probably because companies in both sectors use similar legal language in their descriptions even though they operate in different industries.

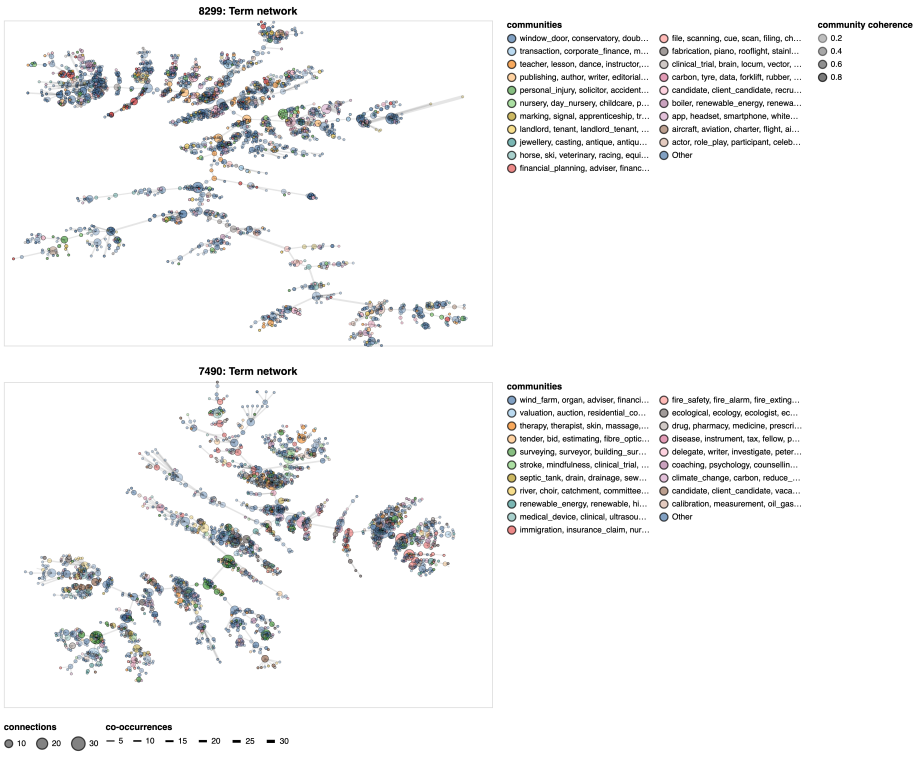


Figure 18: Term networks for a random sample of 1000 companies in the 8299 (Other Personal Services Not Elsewhere Classified) and 7490 (Other professional, scientific and technical activities Not Elsewhere Classified). Each node is a term, its connections indicate co-occurrences in companies and its colour the community (sector) that it belings to. The size of the nodes represent their total occurrences in the sampled corpus and their opacity the coherence of its names.

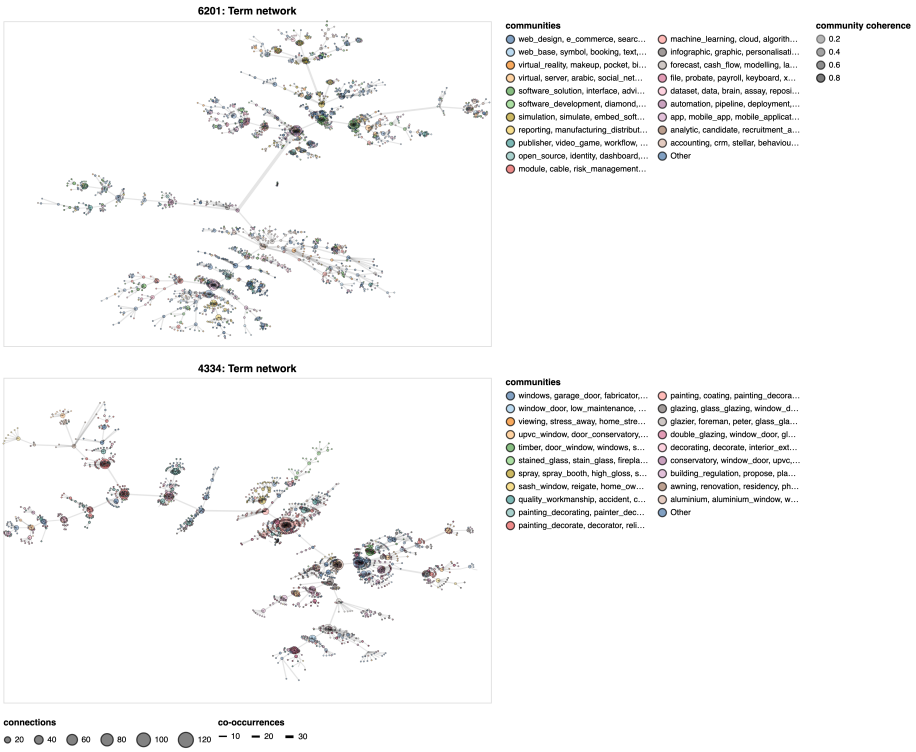


Figure 19: Term networks for a random sample of 1000 companies in the 6201 (Computer Programming Activities) and 4334 (Painting and glazing). Each node is a term, its connections indicate co-occurrences in companies and its colour the community (sector) that it belings to. The size of the nodes represent their total occurrences in the sampled corpus and their opacity the coherence of its names.

Table 10: Examples of duplicated communities

| Community 1 | Community 2 |
|---|---|
| cash_flow, tax, corporate_finance, tax_advice, gap, business_owner, financial_plan, fossil_fuel, guess, potential_buyer | addition_traditional, tax_advice, corporate_finance, auditing_accounting, client_alike, general_accountancy, build_enviable, tax, audit_tax, accounting_tax |
| franchise, dealership, car_dealer, heating_oil, agricultural_machinery, sons, petrol_station, circa, family_tradition, petrol | car_van, stocklist_regularly, sure_update, worth_give, reach_right, vehicle_maintenance, guarantee_receive, suit_budget, sell_car, motorhome |
| tyre, trailer, cruise, commercial_vehicle, coupling, hitch, quantity_warehouse, minibus, tow, brake | extra_mean, mind_showroom, confident_disappoint, car_ready, selection_quality, buy_car, stock_wide, bike, yoga, orbit |
| equity_release, leasing, car_van, lease, vehicle_leasing, contract_hire, vehicle_rental, dealership, insurance_policy, insurance_company | landlord_tenant, legislation_attention, detail_conscientious, law_government, change_tenancy, strive_forefront, landlord_seek, tenant_search, property_letting, local_knowledge |
| ensemble, recording, instrument, guitar, musician, choir, composer, orchestra, concert, piano | extra_mean, mind_showroom, confident_disappoint, car_ready, selection_quality, buy_car, stock_wide, bike, yoga, orbit |

#### 5.4.5 Nearest-sector analysis

Our company labelling based on the size of the overlap between company descriptions and community vocabularies yields around 50,000 labelled companies (less than 15% of our sample). One explanation for this is that our filtered lists of tokens do not fully capture the vocabulary of terms generated by a sector (i.e. $C_k \subset S_k$). In order to address this and improve the recall of our analysis, we implement a semantic similarity stage in our analysis where we:

1. Vectorise all documents (company descriptions) in our corpus using doc2vec (Le and Mikolov 2014),
2. Create vector summaries of each sector (the mean of all companies labelled with that sector) focusing on 582 sectors with at least 50 companies in it
3. We tag each unlabelled company with at least four terms in its description with its closest sector in vector space if the similarity score is above a minimum threshold of 0.75.

This strategy helps us label an additonal group of 88,000 companies. We present some randomly drawn examples in table 10. Although the matching strategy

works well in most cases, we find some exceptions such as a rooftop restaurant matched with a roofing company. False positive matches could be removed by setting a higher threshold for the semantic match at the expense of a loss in recall.

Table 11: Randomly drawn examples based on our semantic similarity expansion.

| Comm | Salient terms | Description of matched company |
|---|---|---|
| 7911_11 | refund, car_rental, car_hire, cancel, cancellation, deposit, reserve_right, booking, non_refundable, liability | If you or any other member of your party wishes to cancel a holiday ( or any part thereof) which is subject of a confirmed booking notice of cancellat... |
| 6209_8 | digital_marketing, web_development, brand_awareness, linguistic, business_acumen, seo, disruptive, territory, secret, digital_channel | Positive Partnerships has been running high-impact digital marketing campaigns for top luxury, travel & lifestyle brands in Europe, Asia and the Ameri... |
| 4299_4 | main_contractor, machining, sub_contractor, estimating, precision_machining, machining_centre, tooling, sub_contract, prototype, turning | We specialise in all aspects of structural timber and steelwork detailing, including wall panels, floor joists and roof trusses. Examples of work unde... |
| 9312_3 | sailing, sail, sailor, yacht, crew, dinghy, vessel, wind, harbour, racing | Elite Sailing offers RYA sail and motor cruising training at Chatham on the Medway in Kent – less than one hour from central London. Elite Sailing is ... |
| 9999_64 | fish, ice, seafood, fresh_fish, fishing, fishery, catering_industry, cold_storage, temperature_control, cold_store | Bioconsortium LTD is an active global trading firm. Our core trading activities are in ecological biomass market. We import & export wood pellets for ... |
| 8690_27 | gym, fitness, personal_trainer, workout, fitness_level, motivation, zone, physical_activity, fitness_class, trainer | We believe in you, we are invested in you. At Fit to the Core we provide up to 20 Pilates classes and everyone is ... |

| Comm | Salient terms | Description of matched company |
|---|---|---|
| 9319_1 | golf, cardinal_hole, golf_course, clubhouse, golfer, hole, golfing, golf_club, tee, par_cardinal | We have a proven track record in coaching golfers of all levels to improve and achieve their goals. You have the opportunity to video your swing, bunk... |
| 8899_4 | mental_health, mental_illness, social_inclusion, service_user, peer, wellbeing, mentoring, seeker, eat, fulfil_potential | We strive to ensure that our classes are non-intimidating and welcoming to ensure full inclusion. We also deliver services to individuals in their own... |
| 8299_32 | wedding_dress, dress, bride, gown, wardrobe, bridal, fancy_dress, bedroom, lounge, alteration | We are proud to stock products that incorporate the latest technology ensuring that they are simple to fit and built to last. Here at Barrier Componen... |
| 8899_4 | mental_health, mental_illness, social_inclusion, service_user, peer, wellbeing, mentoring, seeker, eat, fulfil_potential | TalkLife's vision is to reduce self harm and suicide globally, and make a positive impact on the lives of our users. We believe that utilising the lat... |
| 7490_2 | candidate, client_candidate, vacancy, interview, job_seeker, recruitment_agency, recruiter, recruitment_solution, recruitment_consultancy, accountancy | Clark Wood is a specialist Recruitment Consultancy servicing the Public Practice & Tax sectors operating throughout the United Kingdom. Our core clien... |
| 7311_0 | digital_marketing, medium_relation, digital_strategy, content_marketing, outsource, search_engine, thought_leadership, sme, retention, data_drive | By tapping into the right kind of support and advice, an organisation can be more confident that it will develop more successful marketing strategies ... |
| 4299_4 | main_contractor, machining, sub_contractor, estimating, precision_machining, machining_centre, tooling, sub_contract, prototype, turning | Established since 1989, we are one of the south west's leading Roofing & Leadwork Specialists and mainly carry out our services in Dorset Somerset and... |

| Comm | Salient terms | Description of matched company |
|---|---|---|
| 7112_13 | plumbing_heating, boiler, plumber, boiler_repair, plumber_heating, appliance, wood_pellet, central_heating, plumbing, job_big | In Autumn 2018 Balcas acquired the wood pellet trade and assets of the leading GB wood pellet distributor Forever Fuels Ltd., enabling Balcas to furth... |
| 6820_3 | bedroom, en_suite, upvc_window, bed_breakfast, breakfast, guest_house, dining_room, lounge, bathroom, conservatory | We're Norwich's first and only rooftop restaurant and bar, with 360-degree panoramic views of famous landmarks including the Cathedral, Castle and Car... |
| 4299_4 | main_contractor, machining, sub_contractor, estimating, precision_machining, machining_centre, tooling, sub_contract, prototype, turning | Work for one of the UK's leading manufacturers and distributors for safety and security systems. At Channel, safety and security is in our DNA, we pri... |
| 4299_4 | main_contractor, machining, sub_contractor, estimating, precision_machining, machining_centre, tooling, sub_contract, prototype, turning | Driving success since 2009 we are committed to helping our pupils all stages of driving lessons in Manchester from complete beginners to more advanced... |
| 9609_58 | payroll, accountant, accounting, accountancy_service, sole_trader, tax, taxation, bookkeeping, accountancy_taxation, accounting_software | Inspire works in partnership with professional services firms to understand their business development (BD) challenges and objectives. We provide besp... |
| 4321_6 | control_panel, main_contractor, control_system, electrical_installation, control_automation, increase_turnover, lighting_control, gas_engineer, heating_ventilation, curve | Proctor Enterprises Limited provides electrical contracting services for private and public sector customers, covering a wide range of domestic and co... |

| Comm | Salient terms | Description of matched company |
|---|---|---|
| 7022_31 | renewable_energy, renewable, energy_saving, reduce_energy, solar, biomass, heat_pump, electricity, fossil_fuel, led | At Evidence & Analytics we provide sophisticated analytical solutions and robust research, most appropriate to your business objectives. The team at E... |

Figure 20 shows company counts by community focusing on the top 40 communities by number of businesses, and their SIC4 division. We note that the division distribution for companies suggests a substantial degree of misclassification (e.g. there are many companies in the SIC code for Division M: Professional, Scientific and Technical Activities labelled with the 'hair salon' new sector). This is caused by overlaps in language across sectors and the presence of equivocal terms in our community vocabularies. Future work should focus on strengthening the procedure we use to tighten the overall vocabulary linked to specific sectors, and to create the dataset of labelled companies.

## 5.5 Sketching a bottom-up industrial taxonomy

We conclude our analysis by prototyping a bottom-up industrial taxonomy based on sector co-occurrences in companies. In a similar way to our analysis of terms inside companies, here we assume that sectors that tend to co-occur in the same companies are related to each other i.e. belong to a higher level category. We decompose that network using community detection again and consider what sectors tend to be placed in the same "higher level communities."

Figure 21 and figure 22 respectively show the resulting sector co-occurrence network and the SIC section sources of the sectors assigned into higher level communities (i.e., the Sections containing the SIC4s that we extracted those sectors from) for a randomly selected number of communities (table 12 shows some "central sectors" within those communities).

We note that there are some cases where the high level communities are assigned sectors extracted from the same SIC division while in others they are assigned sectors from a wide range of divisions, suggesting differences between the SIC structure and an industrial taxonomy based on a bottom up analysis of sector co-occurrence in companies. We note that the bottom-up taxonomy seems to cluster sectors based on their market orientation (e.g. high level community 0 seems to be focused on services related to the property market including electricians as well as lawyers) rather than their position within more traditional sectoral segments. This illustrates the potential for using a bottom-up industrial taxonomy to explore the economy from novel angles.

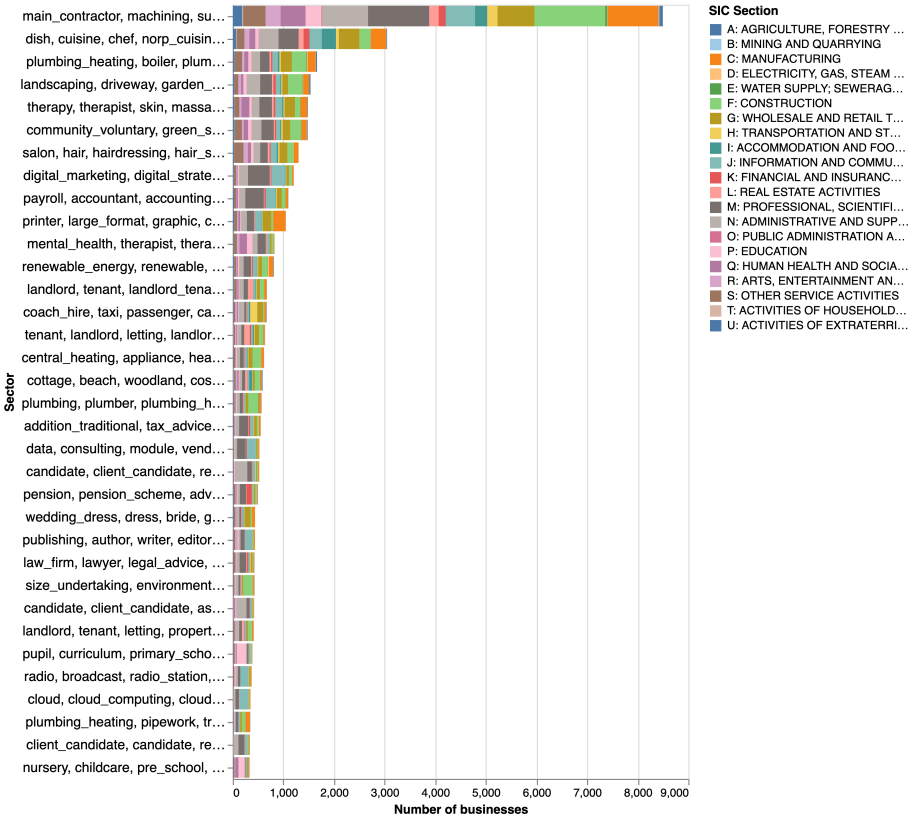


Figure 20: Count of companies by the sector they have been labelled on and their SIC4 division

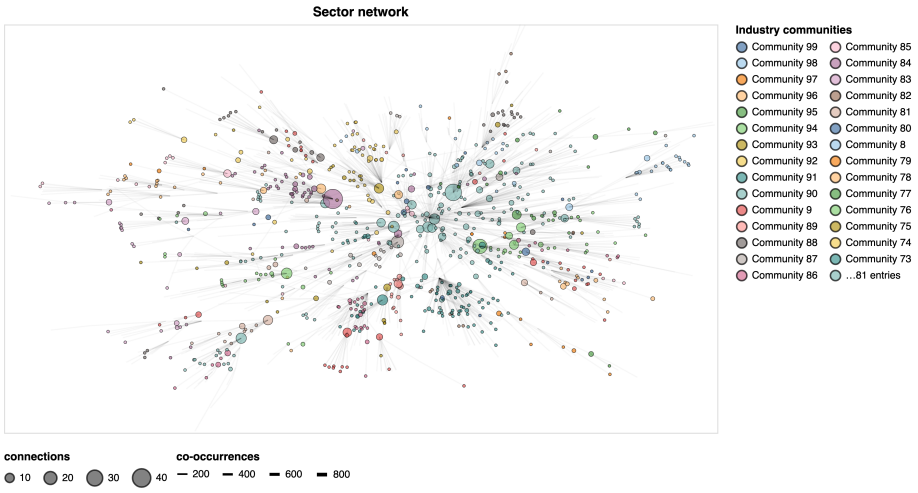


Figure 21: Sector co-occurrence network: each node represents a sector obtained through the process described above. The size of the nodes represent the degree centrality of the node, and its colour the community that it belongs to

Table 12: Central sectors inside high-level communities.

| High level community | Central sectors |
|---|---|
| 0 | electrical_installation, electrician, ordinal_edition, approve_contractor, fault_finding, consumer_unit, wiring, electrical_contractor, testing_inspection, electrical_contracting____legal_advice, lawyer, law_firm, solicitors, solicitor, complaint, legal_issue, approachable, immigration, traditional_value____tenant, tenancy, landlord, let_agent, negotiate, rental_income, appliance, cardinal_bed, end_tenancy, grinding |
| 2 | software_development, diamond, system_integration, bespoke_software, custom_software, software_developer, life_cycle, sage, outsourcing, web_base____carrier, reseller, telecoms, interconnect, termination, traffic, bundle, voice_datum, tier_cardinal, telephony____cyber_security, risk_management, square, forensic, attack, identification, apt, governance, cyber, auditor |

| High level community | Central sectors |
|---|---|
| 5 | motor_trade, finance_package, low_mileage, drive_away, forecourt, prestige_car, mpv, mot, subject_status, train_station____bouncy_castle, inflatable, soft_play, castle, bounce, slide, public_liability, toddler, ball, jungle____home_start, train_volunteer, time_week, recruit_train, co_ordinator, stringent, breastfeed, young_child, confidential, secure_funding |
| 6 | bathroom, tile, bathroom_suite, flooring, bathroom_installation, shower, plumbing, bedroom, wet_room, plumbing_heating____bathroom, bathroom_installation, plumbing_heating, caravan, central_heating, design_installation, malt, boiler, boiler_servicing, heating_system____bathroom, kitchen_bedroom, tile, bathroom_furniture, architectural_interior, finishing_touch, cabinetry, bathroom_fit, bathroom_suite, bathroom_showroom |
| 16 | solicitor, mediation, mediator, dispute, personal_injury, litigation, probate, court, dispute_resolution, compensation____lawyer, law_firm, solicitor, family_law, personal_injury, litigation, solicitors, dispute_resolution, firm_solicitor, probate____rehabilitation, physiotherapy, injury, physiotherapist, mobility, medico_legal, expert_witness, personal_injury, clinic, litigation |
| 18 | skin, skincare, fragrance, anti_ageing, practitioner, skin_type, clinic, hair, non_invasive, salon____wool, bamboo, skin, nutrient, sheep, hardwood_flooring, fleece, mask, moisture, resistant____wipe, skin, feather, owl, formula, nutrient, dust, medical_device, alcohol, wash |
| 19 | real_estate, loan, borrower, lender, transaction, lending, private_equity, institutional_investor, debt, mortgage____broker, insurer, underwriter, real_estate, underwriting, lender, insurance_policy, loan, general_insurance, insurance_company____complaint, consumer_credit, trading_style, mortgage_insurance, authorise_regulate, debt, fairly, dissolve, include_limitation, repayment_mortgage |
| 20 | gallery, fine_art, contemporary_art, art_gallery, painting, collector, museum, sculpture, artwork, auction____time_walk, museum, train_station, bus, en_suite, shop_restaurant, bed_breakfast, promenade, bar_restaurant, world_famous____auction, house_clearance, painting, valuation, silver_jewellery, fine_art, auction_house, career_span, stamp, buyer_seller |

| High level community | Central sectors |
|---|---|
| 25 | renewable_energy, turbine, renewable, wind_turbine, green_energy, biomass_boiler, project_manager, electrician, renewable_technology, boiler____renewable_energy, renewable, energy_saving, reduce_energy, solar, biomass, heat_pump, electricity, fossil_fuel, led____solar, wireless, temperature, controller, electricity, relay, gauge, harvest, sensor, battery |
| 34 | green_deal, emission, energy_saving, carbon_footprint, flue, greenhouse_gas, reduce_carbon, refinery, energy_usage, climate_change____display_showroom, marble_granite, post_war, establish_enviable, gas_electric, modern_contemporary, cast_iron, plastering, insert, fireplace |
| 40 | carer, carers, family_carer, disabled_people, nurse, disabled, parent_carer, information_advice, illness, homecare____carer, family_carer, personal_datum, caregiver, service_user, companionship, maintain_independence, elderly, live_independently, medication____advocacy, disability, learn_disability, advocate, window_film, fair_trade, people_learn, learning_disability, dementia, people_live |
| 46 | t_shirt, garment, embroidery, martial_art, shirt, paint_colour, cardinal_t, colourful, uniform, candy____reinforce_concrete, design_construction, conveyor, structural_engineering, chimney, mobile_crane, flexible_adaptable, civil_structural, sons, building_civil____staircase, stair, joinery, hardwood, loft_conversion, site_survey, shop_front, flight, period_property, skilled_tradesman |
| 53 | camera, photography, photographer, shot, high_resolution, sensor, footage, detection, predecessor, photographic____collar, battery, cardinal_g, cat, signal, satellite, compatibility, detect, leak_detection, antenna____animation, film_video, technical_ability, graphic_design, visual_communication, shortcut, cgi, recently_join, logo_design, video_production |
| 61 | service_user, health_social, service_users, old_people, person_centre, carer, shall, nursing_home, live_independently, promote_independence____service_user, person_centre, health_social, medication, independent_living, mental_health, adult_learn, hospice, companionship, trainee |
| 67 | pension, workplace_pension, auto_enrolment, pension_scheme, retirement, pension_fund, tax, mis_sell, threshold, self_employ____debt, retirement, pension, learner, borough, reliant, trustee, pension_scheme, education_training, statutory____retirement, retirement_planning, insurer, income_protection, financial_planning, simon, investment_portfolio, critical_illness, mortgage, cat_dog |

| High level community | Central sectors |
|---|---|
| 72 | transaction, corporate_finance, capital_market, m&a, emerge_market, advisory_firm, mining, institutional_investor, car_leasing, valve____transaction, corporate_finance, m&a, cross_border, advisory_firm, financial_advisory, private_equity, brokerage, investment_bank, financing |
| 84 | meat, butcher, pork, goat, animal_welfare, rear, food_mile, burger, breed, sausage____meat, sausage, butcher, butchery, alliance, pork, table_tennis, net, lamb, cheese |
| 89 | trailer, vehicle_trailer, low_loader, ramp, semi, truck, heavy_plant, train_operative, refrigerant, truck_trailer |
| 90 | ice_cream, kiosk, cream, fruit, tub, flavour, cart, farm_shop, chocolate, café |
| 91 | asbestos, asbestos_removal, asbestos_survey, asbestos_management, licensed, risk_assessment, hse, control_panel, remediation, waste_collection |
| 93 | corporate_governance, auditor, risk_management, non_executive, remuneration, effectiveness, committee, financial_reporting, chairman, executive_director |
| 94 | cottage, golf_course, short_walk, historic_city, sea_view, acre, function_room, storey, manor, equipped____cottage, self_cater, sightseeing, hostel, picturesque_village, self_catering, amenity, wildlife, holidaymaker, city_break |
| 98 | self_storage, container, storage_unit, shipping_container, van_hire, simplify, storage_facility, cardinal_ft, depot, storage_space____self_storage, storage_facility, container, secure_storage, cardinal_ft, storage_container, relocation, storage_solution, belonging, file |
| 105 | home_automation, audio_visual, audio, digital_signage, climate_control, conference_facility, conference_room, music_video, sensory, ease_use |
| 106 | gym, fitness, personal_trainer, workout, fitness_level, motivation, zone, physical_activity, fitness_class, trainer |
| 109 | ecological, ecology, ecologist, ecological_consultancy, bat, protect_specie, mitigation, planning_application, biodiversity, planning_authority____reptile, habitat, wildlife, protect_specie, crop, cocoa, specie, ecologist, illegal, fat |

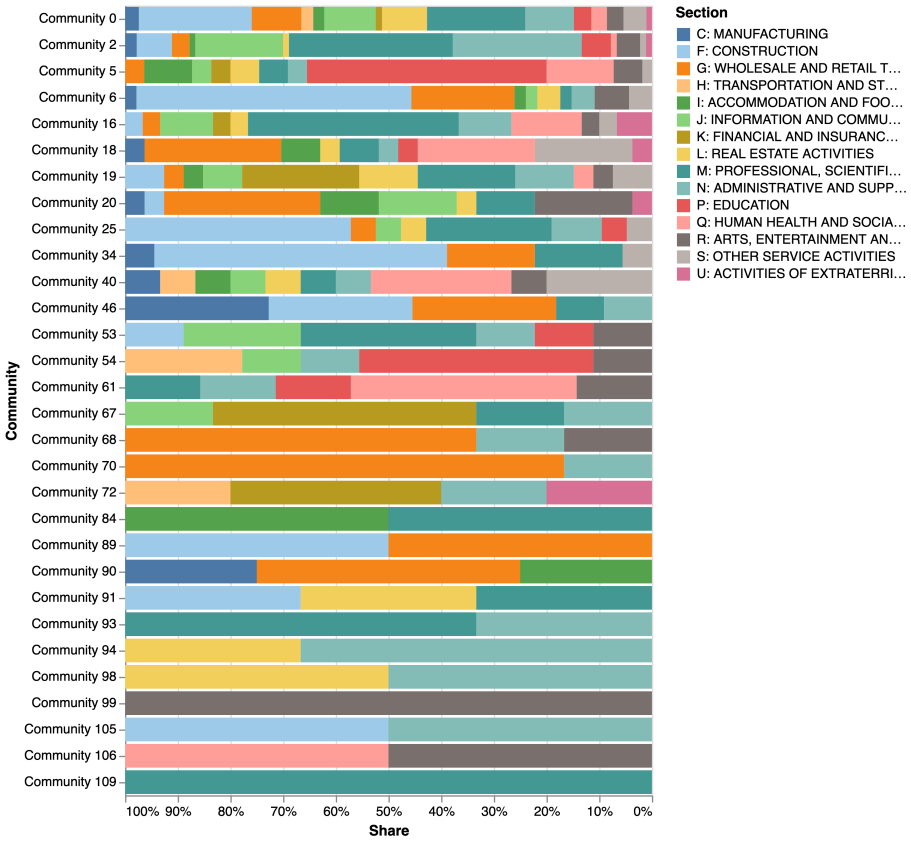


Figure 22: Community composition: Distribution of section sources for sectors assigned to different high-level communities

## 5.6 Conclusions and next steps

The proof-of-concept method to build an industrial taxonomy "from the bottom up" that we have presented in this section illustrates the opportunities and challenges for using novel data sources and data science methods to study the industrial composition of the economy with more granularity and timeliness than is possible using official taxonomies: our approach is able to extract highly granular sectors from heterogeneous and uninformative SIC4 codes, including new activities related to green and sustainability agendas, industries working with emerging technologies, and digitalised segments of existing industries (e.g. "digital marketing").

Unfortunately, this requires building a complex pre-processing, community extraction, labelling and post-processing pipeline with multiple steps and tunign parametres. The quality of the outcomes of our final step to classify companies into the new sectotal categories still leaves much to be desired, providing a focus area for future work.

There are several areas for improvement in our pipeline. They include:

- Exploring alternative strategies to generate our term communities: Some of the challenges that we face when labelling companies with our new industry categories appear to stem from imbalances in the number of terms assigned to different communities at the community extraction stage, and to the generic nature of some of these terms. Some potential strategies to address this include implementing community detection methods that allow overlaps between communities so that fuzzy terms are not crisply assigned to a single community in a way that could skew our results, and weighting the presence of a sector in a company by the salience that different terms have in it.
- Exploring alternative strategies to generate our labelled sector groups: Another option would be to replace the community extraction step and term-based labelling with an alternative strategy where we cluster documents (companies) based on the semantic similarity of their descriptions. We would then identify salient terms in these groups to generate sector labels and classify other companies into those sectors based on document similarity. This could drastically simplify the pipeline we have adopted here.
- Expanding and enhancing the bottom-up taxonomy: We have only began exploring the potential for generating a bottom-up taxonomy based on sector co-occurrences in companies. A next step, to be taken after we have improved the robustness of our company labelling approach, would be to search for an optimal partition of the resulting network, and to develop an industrial hierarchy where sectors are agglomerated at various levels of resolution, mimicking the 'section / division / class / subclass' structure present in the SIC taxonomy.
- Evaluating the robustness of results to different parametres: Our pipeline

includes a host of parametres that we have selected heuristically after a qualitative exploration of outputs. A next step is to formalise this testing and tune our parametres to improve performance.

# 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 a network analysis of term co-occurrence in company descriptions: new ecomomic activities - for example around renewable energies and emerging technologies - 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, for example around sustainability. 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 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 multiple processing steps and a great degree of parameter tuning which at this pilot stage we have undertaken empirically. Our decision to focus on sectors with high levels of activity in the data leads to the removal of SIC codes in primary sectors that are less well-covered by the glass data. Perhaps most worryingly, the protocol that we have developed in order to assign companies to sectors, a critical step before we are able to deploy the taxonomy to yield economic statistics generates a substantial number of erroneous matches. All 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 sector such as for example renewable energies 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, this 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.

Asano, Tetsuo, Binay Bhattacharya, Mark Keil, and Frances Yao. 1988. "Clustering Algorithms Based on Minimum and Maximum Spanning Trees." In *Proceedings of the Fourth Annual Symposium on Computational Geometry*, 252–57.

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 (1): 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.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal*

*of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.

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.

Enright, Anton J, Stijn Van Dongen, and Christos A Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." *Nucleic Acids Research* 30 (7): 1575–84.

Fortunato, Santo, and Darko Hric. 2016. "Community Detection in Networks: A User Guide." *Physics Reports* 659: 1–44.

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.

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.

Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 1.

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.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *International Conference on Machine Learning*, 1188–96. PMLR.

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

Loper, Edward, and Steven Bird. 2002. "Nltk: The Natural Language Toolkit." *arXiv Preprint Cs/0205028*.

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."

McDiarmid, Colin, and Fiona Skerman. 2018. "Modularity of Erdős-Rényi Random Graphs." https://arxiv.org/abs/1808.02243.

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

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–51.

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.

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. 2007. "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks." *Physical Review E* 76 (3): 036106.

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

Rossetti, Giulio, Letizia Milli, and Rémy Cazabet. 2019. "CDLIB: A Python Library to Extract, Compare and Evaluate Communities from Complex Networks." *Applied Network Science* 4 (1): 1–26.

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*.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-41695-z.

Ustalov, Dmitry, Alexander Panchenko, Chris Biemann, and Simone Paolo Ponzetto. 2019. "Watset: Local-Global Graph Clustering with Applications in Sense and Frame Induction." *Computational Linguistics* 45 (3): 423–79.

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.

Zhang, Wei, Xiaogang Wang, Deli Zhao, and Xiaoou Tang. 2012. "Graph Degree Linkage: Agglomerative Clustering on a Directed Graph." In *Computer Vision – ECCV 2012*, edited by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, 428–41. Berlin, Heidelberg: Springer Berlin Heidelberg.