Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Transforming survey analysis: tools for central banks[1]

Nicholas Gray and Dominic Jones,
Reserve Bank of Australia

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Transforming survey analysis

Tools for central banks

Nicholas Gray and Dominic Jones[1]

Abstract

We provide recommendations based on learnings from survey exploratory text analysis at the Reserve Bank of Australia. We provide recommendations around a broad framework of *gradual supervision*, *exploration and iterative refinement*, and *modelling co-occurrences*.

The views expressed in this paper are those of the authors and should not be attributed to the Reserve Bank of Australia. Any errors are the sole responsibility of the authors.

Keywords: text analysis, survey analysis, natural language processing, weakly supervised, text classification.

JEL classification: C80, C83, C55, C45, C19

[1] Reserve Bank of Australia

Motivation

Surveys are commonly used at central banks. For example, surveys on consumer payments, trust, banknote design, inflation expectations, or business conditions.

Surveys often contain questions with free-text response. Sometimes they are used to complete incomplete lists (e.g. "if other, please describe…"), but frequently they are open-ended. For example, a question that asks a respondent about their views on holding cash. These questions are useful for eliciting what's "top of mind" [2]. The downside is that free-text responses are unstructured – they cannot easily be graphed, modelled, or otherwise analysed. Some sort of transformation needs to occur first. In our experience, this often means "coding": extracting a set of key themes and tagging each response with the relevant ones. In the language of text analysis, this is a classification task.

If the survey sample is small enough, manually reading the responses is the best option. In many cases, the sample is large – on the order of thousands of responses, sometimes more. Often, the burden of coding so many responses is too high. However, modern NLP techniques can help.

Framework

In our experience, analysing surveys (or other text corpora) follows a few key themes. We have written them down as a framework to share our learnings and contribute to discussion in the central banking community. The framework has three parts. The parts are not mutually exclusive, but we think they are usefully considered separately.

The first part of the framework is *gradual supervision*. Subject-matter experts often learn throughout the analysis process, and it's important to keep the analysis up to date with the evolving understanding. We make heavy use of 'weakly-supervised' classification techniques here and close collaboration with subject-matter experts.

The second part is *exploration and iterative refinement*. Here, we provide tools and recommendations that help subject-matter experts to explore and comprehend the analysis at any stage.

The third and final part is *modelling co-occurrences*. We are often asked about how key themes co-occur. It has been useful for us to formally model this in many analyses. We provide detail and recommendations about how to do it.

Gradual supervision

Usually, subject-matter experts start knowing perhaps a little – a prior, or maybe some analysis of a subsample. As analysis proceeds, we learn more about the subject. Ideally, the analysis itself should be responsive to this growing knowledge.

Traditionally, NLP techniques are grouped under two headings: unsupervised and fully supervised. Unsupervised techniques take nothing or very little as input, and therefore are easy to use, but can be difficult to interpret. Fully supervised techniques take expert-annotated 'gold' data as input, and allow for state-of-the-art classifications, but are expensive. Both are, of course, useful. Recently, 'weakly supervised' techniques have become more powerful and common, bridging the gap between un- and fully-supervised methods. We now use these methods to keep up with the growing understanding – and thus supervisory power – resulting from analysis. In this section, we will briefly discuss the methods we use, why, and conclude with some recommendations.

Unsupervised

Descriptive statistics like term or phrase frequencies are easy to run and tend to be interpretable. In jargon-rich fields, term frequencies tend to be more useful. We often pair word-level frequencies with phrase-level frequencies. Phrases – several words better understood as a single unit of meaning – can be extracted in several ways. The simplest is to compute n-gram frequencies, but more sophisticated phrasal segmentation algorithms exist.

Unsupervised topic models – we tend to use BERTopic [1] – are workhorses in this category. Rather than term frequencies, topic models extract 'topics' from a set of documents. Often, these topics are a set of terms, and a document can be represented by a mixture of topics. This maps neatly on to questions usually asked at early stages of an analysis: "what sort of things are people saying?"

Plotting "embedding spaces" and iteratively refining them is a common technique in this category, discussed below.

Weakly supervised

Weakly supervised approaches fill in the gap between unsupervised and fully supervised methods. When we don't have labelled data, these methods allow us to better use whatever guidance we *do* have – be it a list of queries, keywords, examples, or a taxonomy.

Weakly supervised methods have much similarity with information retrieval methods. There is some degree of supervision given to the algorithm – which might otherwise be called a *query*. We then ask what documents are relevant to the query. Therefore, traditional information retrieval methods can be useful here whenever the question can be cast appropriately. For example: "which responses are relevant to unemployment?" We frequently use keyword-based relevance schemes (e.g. BM25 [6]) and transformer-based retrievers (e.g. sentence transformers [5]) for this task.

Topic models can also be used in a weakly supervised fashion by providing a list of "seed" words to build topics around. CatE [4] or seeded BERTopic (or many others) can be used for this purpose. These techniques are especially useful to enrich current understandings.

Finally, often survey responses are "coded" into several categories defined by some codebook. For example, a response might be about unemployment and future

expectations but not about investment intentions. While the construction of the codebook itself is mostly appropriate for subject-matter experts, the actual coding is a *classification* task.

Since constructing a fully supervised classifier is often expensive, weakly supervised classifiers are a useful alternative. These may take the form of a dataset seeded by information retrieval techniques, a prompt-based LLM approach, or specialised algorithms like X-Class [8]. While the performance of a weakly supervised classifier is lower than that of a fully supervised one, it is often not too much lower while being much cheaper to put together. We have found this to be important for speeding up feedback loops between data scientists and subject-matter experts.

Fully supervised

That is not to say fully supervised classifiers are obsolete. When circumstance demands the best possible performance, or when the survey is repeated and temporal consistency is useful, traditional fully supervised classifiers are effective. We often reach for a fine-tuned RoBERTa [10] (though we have been experimenting with ModernBERT [9]).

Recommendations for gradual supervision

Work closely with the subject-matter experts. Communicate the nature of what you are doing and be honest about limitations. When you are eliciting guidance – for example, lists of relevant keywords or even a taxonomy – be clear about what format you want it in. Even consider making a form for it.

Reporting matters: spend some time making your reports digestible.

Finally, here are a few questions we tend to ask experts when we're thinking about moving up the chain towards more supervised techniques:

- Is this worth more investigation?
- Should we cut this out of the analysis?
- Are there any indicators of this sort of response – textual or otherwise?
- Are there any sub-categories you might be interested in?
- Would you like to track this over time, if the survey is repeated?

Exploration and iterative refinement

Document clustering is a useful exploratory technique. The most basic form of exploration is simply looking through responses manually. Using spreadsheets can be unwieldly, especially if there are many similar responses. Instead, we often use "embedding spaces". Documents – survey responses – are first converted into numerical "embeddings", which are then plotted in an "embedding space", with potentially some sort of clustering algorithm applied. It's often common to colour points in visualisations of these embedding spaces according to however they've

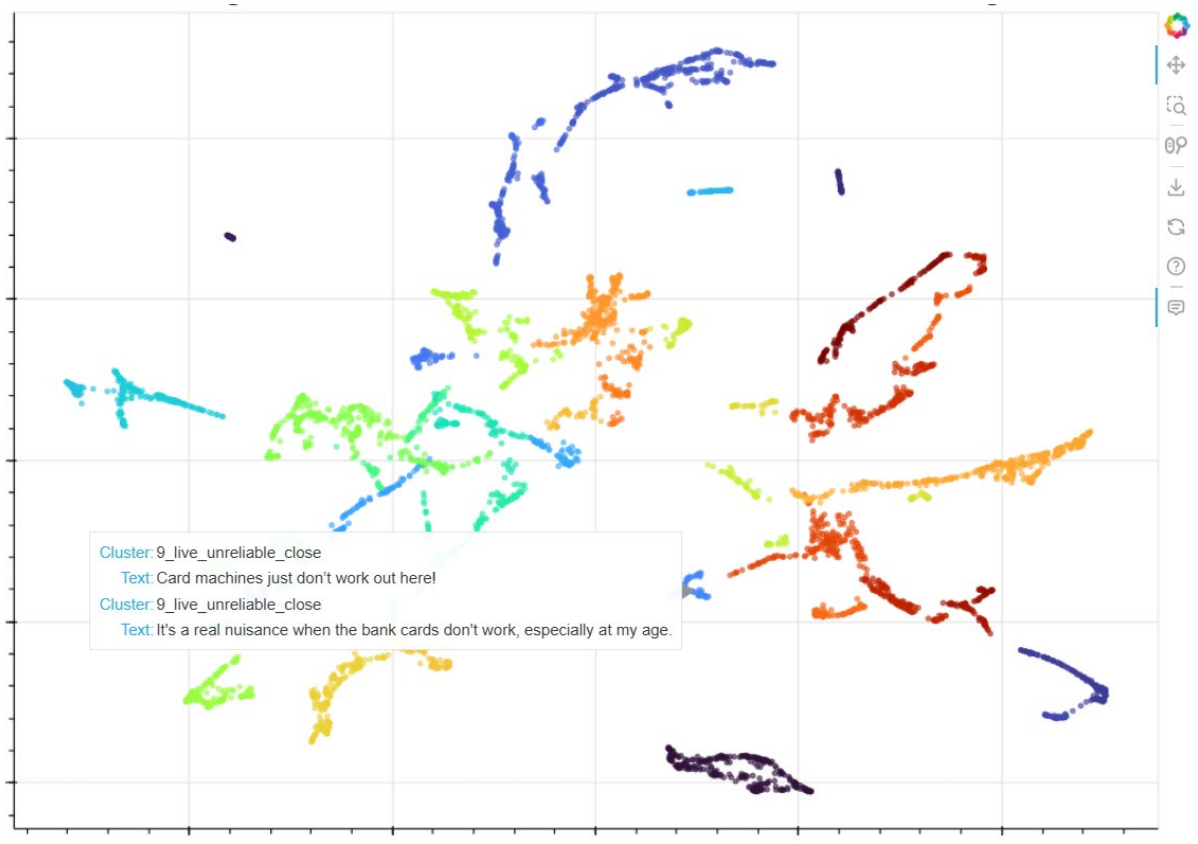
been categorised (e.g. via a gradual supervision pipeline). It's also common to use the embedding space *itself* to recover categories. When working with these embedding spaces we often think of two interrelated steps – *exploration* and *refinement*.

Exploration

Exploring is crucial for building a better understanding of themes in the data. It also makes it easier to explain themes to or loop in subject-matter experts, helping the refinement step. Crucial aids for exploration are *interactivity* and *summarisation*.

2D Visualisation of Sentence Transformer Embedding Space

Dimension reduction with PaCMAP; Clustering with HDBSCAN; Artificial data — Graph 1



Data is artificially generated using ChatGPT to simulate expected responses in a hypothetical "Cash Preference" survey.[2] We used the default settings for PaCMAP.

The user is currently hovering over responses assigned to "9_live_unreliable_close" cluster.

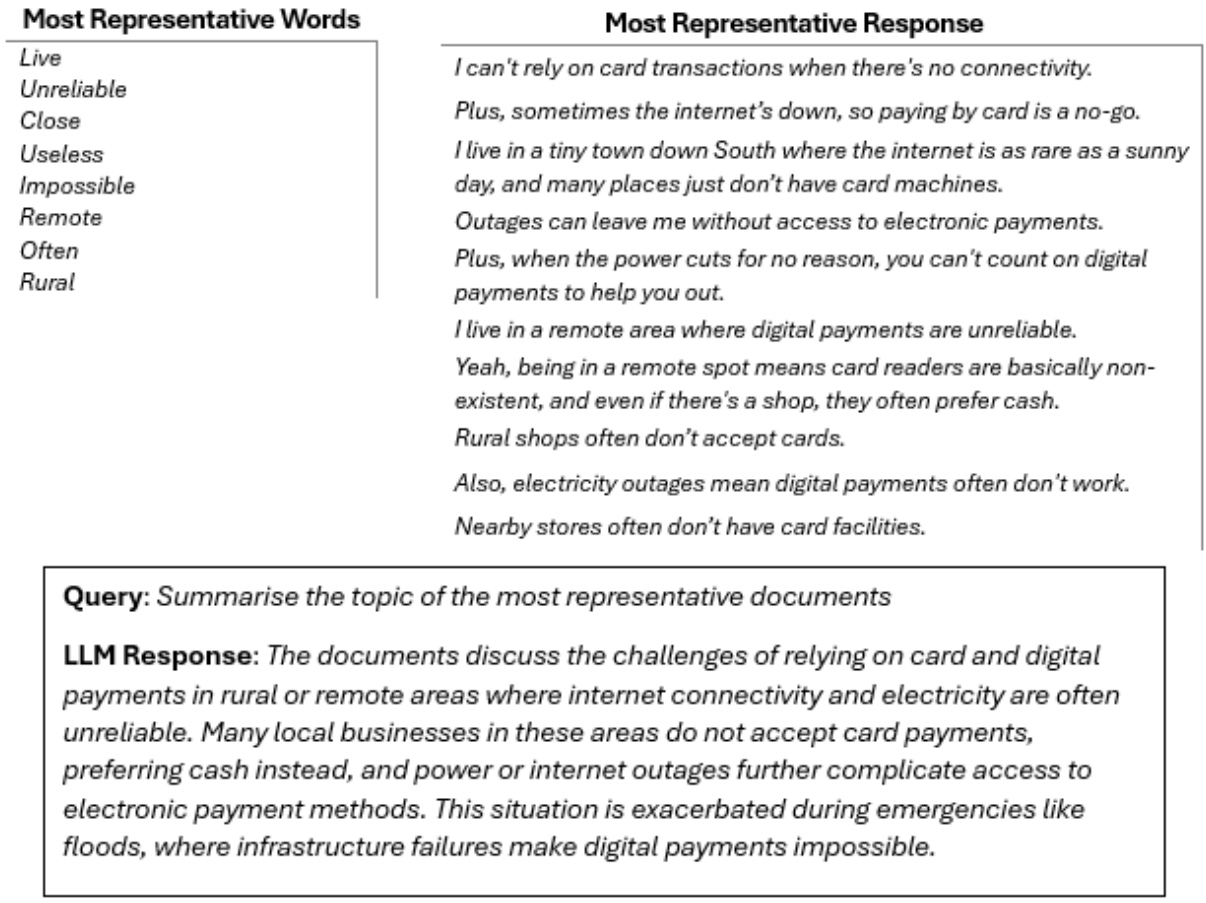
Sources: RBA

[2] All responses are created by the model based on prompts for generalised themes expected in such a survey. No real survey data is provided to the model.

Interactivity. The simplest technique is to apply a dimensionality reduction algorithm such as UMAP [3] or PaCMAP [7] to reduce the embedding space to 2 dimensions. This allows us to draw a scatter plot, where each point represents a document (survey response). The distance between points is proportional to the similarity (however defined) between documents. We can also colour the points based on some categorisation, such as cluster assignments. Interactivity allows users to zoom in on clusters and hover over points to read the text that each point represents. Visually exploring responses to surveys can reveal interesting characteristics of responses and clusters, which can help refine analysis.

Summarisation of Response Cluster

Characteristics for cluster "9_live_unreliable_close"; Artificial data — Graph 2



Same artificially generated data as Graph 1.

LLM summary using locally hosted model[3]

Sources: RBA

[3] Microsoft's Phi-4 model.

Summarisation. Hovering and reading clustered responses one-by-one can only take one so far. To speed up feedback loops, we often *summarise* clusters of responses. For this, we use cluster-indicative keywords, representative responses, and LLM-based summaries.

For cluster-indicative keywords, we often use "class-based" TF-IDF. It extracts words or phrases that are more informative to one class (cluster) than others by weighting term frequency and how common that term is across different clusters. High scoring terms are more representative of responses in the cluster. Informative keywords identified here can also be fed into a weakly supervised classifier or other method to refine response clusters.

For representative responses, we use the geometry of the clusters themselves. Since they live in some embedding space, we can retrieve responses close to the centroid of the cluster. This gives us specific responses that are more representative of the cluster. Reading these responses can provide a quicker and clearer overview of the main themes. Conversely, we can look at the most unusual responses in the cluster by extracting the responses furthest from the cluster centroid. Often these responses are not typical to the central theme of the cluster. These responses can be informative for subject-matter experts or may signal that the cluster needs refinement. By looking at responses from across the cluster geometry, we aim to avoid biases that arise from only considering measures of central tendency.

Finally, we sometimes use (small, local) LLMs to generate summaries of clusters. These are also effective at providing quick and clear summaries of clusters. Depending on how sophisticated the LLM or prompt is, we can also ask it to extract information: for example, keywords, entities. We may even ask it specific questions about responses. If there are token limits, we often use responses near the centroid as above, or even just a random sample.

(Iterative) refinement

Improvements are always identified in whatever preliminary clustering we look at. This is often the step where techniques from gradual refinement are useful. In any case, we may want to combine, split, throw away, or even completely refresh clusters. By gaining insight from exploration techniques and then updating the analysis using appropriate techniques closes the exploration-refinement loop.

The interactive plot, despite being a highly summarised version of the embedding space, can quickly reveal clusters in need of refinement (such as splitting or merging). Looking at the clusters summarised by keywords or representative examples can also indicate whether clusters need further analysis. Finally, using LLMs to summarise clusters speeds up review, especially if there are many.

Recommendations

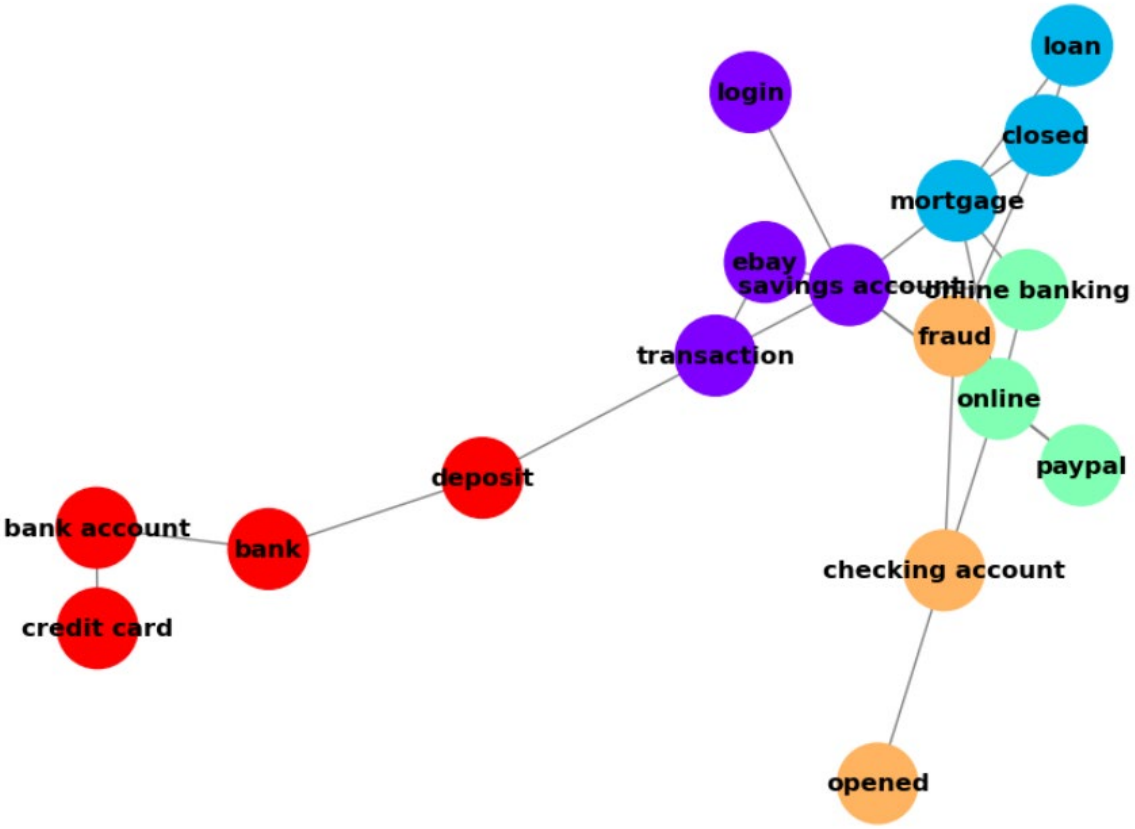
Sit down with your subject-matter experts and explain your charts. Let them see the data underlying any LLM-generated summary. And, where possible, automate the process (e.g. generating summaries or keywords) to tighten feedback loops. Feed refined clusters back into the gradual supervision pipeline – this is often most useful between the un- and weakly-supervised phases.

Modelling co-occurrences

Once we've classified, tagged, coded, or otherwise categorised a set of responses, we're often asked about how categories co-occur. For example, what is the probability of a response being about monetary policy, *conditional* on it being about inflation? Or are there groups of categories that often appear together? Answers to these sorts of questions are often useful for subject-matter experts. There are two ways that we often think about answering these questions. We'll refer to the categories as *concepts* for generality.

Example Concept Co-Occurrence Network

Graph 3



Edges between topics indicate these two topics often co-occur. This was generated from the publicly available CFPB Consumer Complaint Database.

Sources: CFPB, RBA

Bag-of-concepts

A natural way to think about concept co-occurrence is by analogy to topic models. In a simple scenario, a document is a "bag" of words (i.e. the order doesn't matter), and topics are words that often co-occur. One way to estimate these topics is by representing the document bags-of-words as a document-word matrix and then factorising this matrix. Principal Component Analysis or Non-negative Matrix Factorisation are appropriate here. For concept co-occurrence, we have a document-concept matrix (i.e. the output of your classifiers), with each row (column) being a document and each column (row) recording which categories it belongs to. We then factorise in the same way as a simple topic model. The resultant output will highlight groups of concepts that co-occur.

Concept co-occurrence networks

An alternative (but not unrelated) method of modelling co-occurrences is to use tools from network science. For each pair of concepts, we can define conditional probabilities or Jaccard similarity (intersection over union). This is naturally represented as an adjacency matrix, which is interpreted as a graph. From there, we might use filtering heuristics (e.g. we are only interested in links with this much weight) or "backboning" algorithms to filter noise from the network. To find groups of concepts that co-occur, we then apply community detection algorithms to the filtered network. We find this effective as a visual device – network plots can be intuitive once explained. (See graph 3.)

Conclusions

Based on our recent experiences in exploratory analysis of textual survey responses, we present a small framework based around *gradual supervision, exploration and iterative refinement,* and *modelling co-occurrences*. Whatever part of the framework we're thinking about – our general recommendations apply:

- Work closely with your subject-matter experts.
- Use the whole gamut of techniques from un- to fully-supervised.
- Think hard about how to speed up exploration of the data (e.g. via interactivity and summarisation).
- Build tight interactive loops with subject-matter experts to refine your analysis (e.g. cluster groups).
- Use LLMs to reduce cognitive load on experts, but ensure they have access to the underlying data.
- Think about formally modelling co-occurrences if there are a lot of potential overlaps.

References

[1] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794

[2] Ingar K. Haaland, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2024. Understanding Economic Behavior Using Open-ended Survey Data. https://doi.org/10.3386/w32421

[3] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (September 2018), 861. https://doi.org/10.21105/joss.00861

[4] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative Topic Mining via Category-Name Guided Text Embedding. In *Proceedings of The Web Conference 2020*, April 20, 2020. ACM, Taipei Taiwan, 2121–2132. https://doi.org/10.1145/3366423.3380278

[5] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

[6] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *FNT in Information Retrieval* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[7] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. 2021. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *Journal of Machine Learning Research* 22, 201 (2021), 1–73. Retrieved April 30, 2025 from http://jmlr.org/papers/v22/20-1061.html

[8] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021. Association for Computational Linguistics, Online, 3043–3053. https://doi.org/10.18653/v1/2021.naacl-main.242

[9] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. https://doi.org/10.48550/arXiv.2412.13663

[10] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, August 2021. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. Retrieved April 30, 2025 from https://aclanthology.org/2021.ccl-1.108/

RESERVE BANK OF AUSTRALIA

Transforming Survey Analysis

Tools for Central Banks

Nicholas Gray and Dominic Jones
2025 BIS IFC Workshop on Data Science in Central Banking
February 2025

Surveys

- Surveys often include free-text open-ended questions.
  - They’re great for eliciting what’s top of mind – or getting at unstructured thoughts or things the survey had missed.

- If the survey sample is small enough, reading responses is the best option.

- But usually, the sample is huge! NLP techniques can help build a qualitative understanding using quantitative techniques.

Surveys at central banks

- Consumer payments surveys
- Trust and credibility
- Banknote design
- Inflation expectations
- Business surveys
- Lending surveys

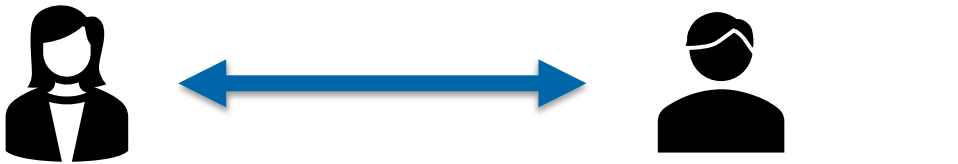
Framework

- Gradual supervision
  - Making use of expert knowledge where we have it…
  - …providing guides where we don't.
- Cluster exploration
  - Providing richer descriptions of topics to help with comprehension
- Iterative refinement
  - Subject-matter experts in the loop.
- Concept matrices
  - Are there any complex interactions worth thinking about?

Gradual supervision

- Usually, SMEs start knowing not much (but not zero) and gradually learn more about survey responses.
- Our old model – just run a topic model – didn't reflect this growing understanding of the data.

- Luckily, techniques now exist that cover the whole gamut of unsupervised to fully supervised – and allow us to incorporate growing expert knowledge!



Unsupervised

Weakly supervised

Fully supervised

- Input: not much
- Easy
- Hard to interpret sometimes

- Input: anything you can think of
- Cheap
- Very good quality

- Input: labelled examples
- Costly + high expert effort often required
- High quality

Unsupervised techniques

- Descriptive statistics: wordclouds, TF-IDF scores, phrasal segmentation algorithms.

- (Unsupervised) Topic modelling
  - BERTopic

- Information retrieval
  - BM25, pretrained sentence transformers.
  - Great for incorporating priors!

Weak supervision

- When SMEs start building a picture it helps to iterate fast.
  - Full supervision is expensive and slows down the loop.
  - Trying candidate models should be cheap.

- Guided topic models
  - BERTopic (again)
  - CatE
- Prompt-guided embeddings
  - Using natural language to shape the embedding space.
- Weakly-supervised classifiers
  - Fast to build!

Full supervision

- When SMEs have a strong understanding of the data – or want to be consistent over multiple surveys – it pays to train a fully supervised classifier.

- We often use **taxonomy classification** and train cross-encoders or text matching networks.

- It’s useful to think about how to communicate these models and to track performance over time.

Recommendations

- Work closely with your SMEs.
  - Communicate the nature of what you're doing and be honest about limitations.
  - When eliciting guidance be clear about what format you want it in. Consider making a form for it!

- Reporting matters.
  - Do spend some time making reports easily digestible.

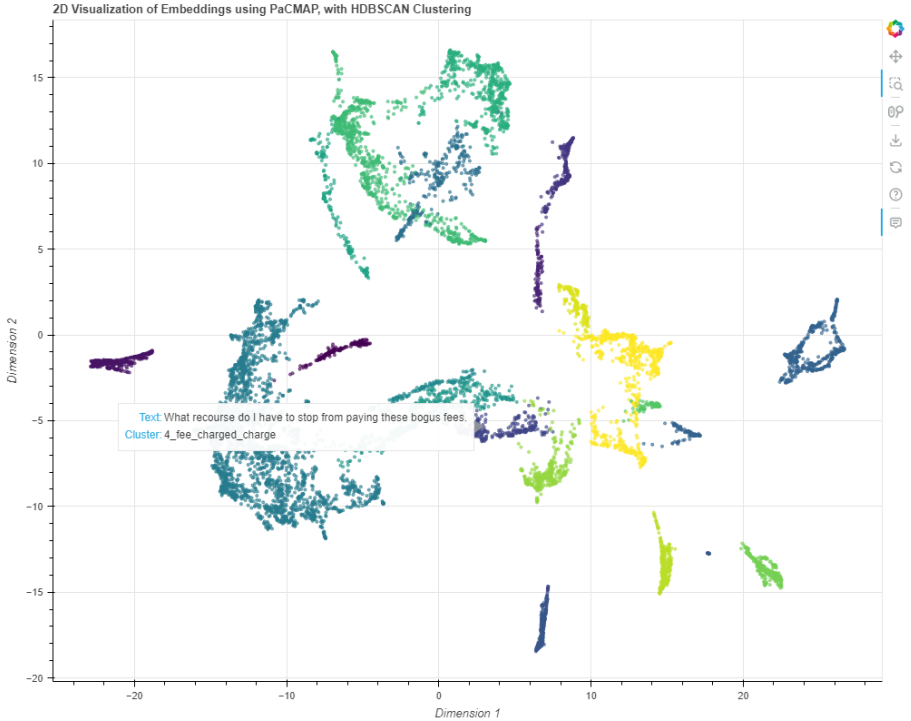
- When should we move up the chain? Questions to ask SMEs:
  - "Is this worth more investigation?"
  - "Should we cut this out of the analysis?"
  - "Are there any indicators of this sort of response?"
  - "Are there any sub-categories you're interested in?"
  - "Do you want to track this over time for the next survey?"

Cluster exploration

- Clusters, groups, classes don’t need to be set in stone.
- Especially when understandings change!
- Agnostic of method, geometry of embedding space can be utilised.
- Providing interactivity and multiple perspectives important for reducing potential bias.

Cluster exploration

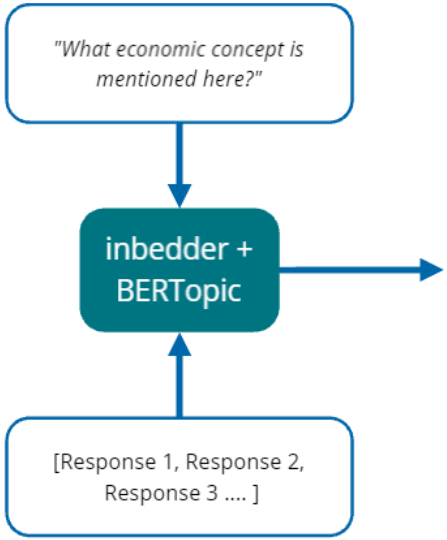
Interactivity:

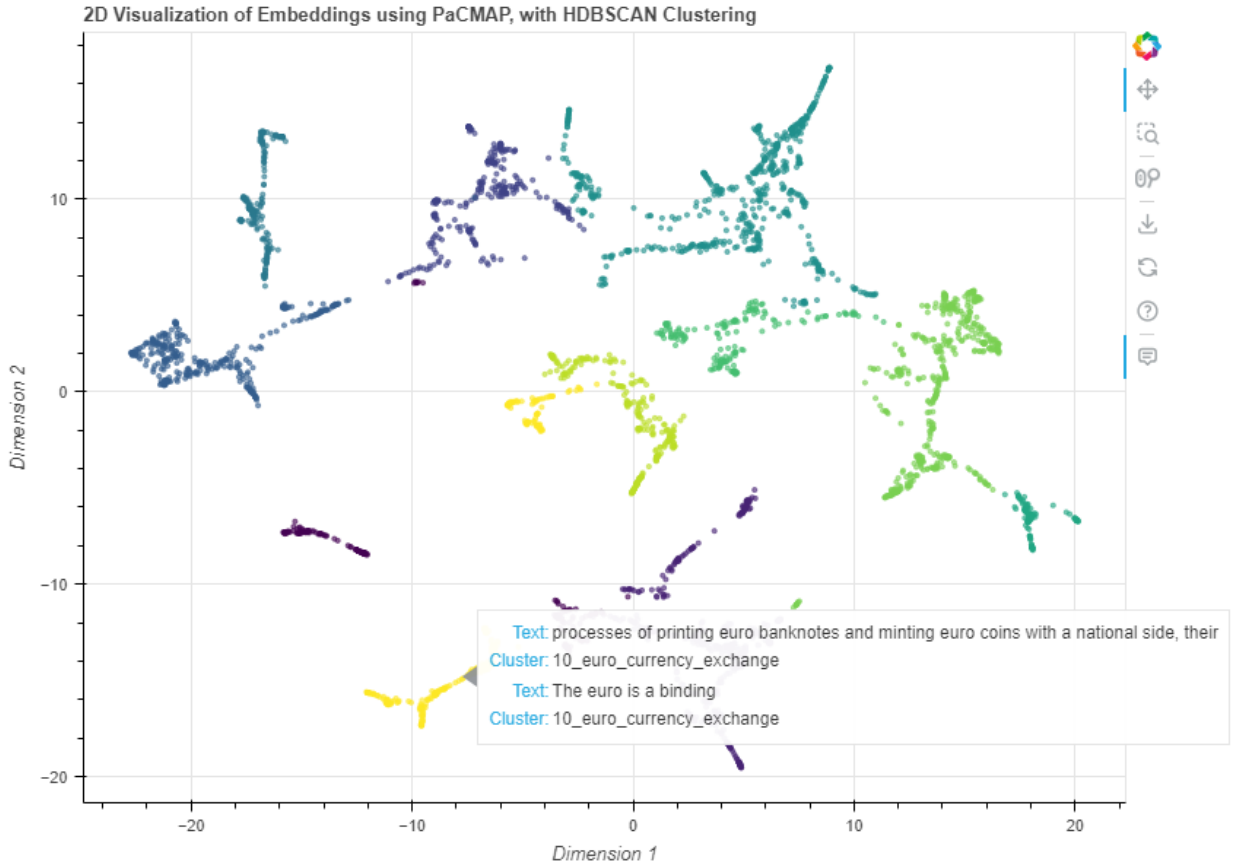


Representative responses:

Closest to centroid of *7_credit_score_report* cluster:	
Distance from centroid	**Reponses**
0.856	I didn't know this was being reported to the credit company until after it was done, I am trying to buy a house and this is causing problems.
0.827	They reported my account to the credit bureaus as 30 days late.
0.813	I contacted my bank in regards to this issue as they are reporting every month to the credit reporting agency 's with remarks which is affecting my credit score in a negative way.
0.805	I told them that was not fair because I cleared up my credit and it is in good standing.
0.804	This company has pulled my credit multiple times for no reason as well.

Prompts, too!


"What economic concept is mentioned here?"
inbedder + BERTopic
[Response 1, Response 2, Response 3 ....]


2D Visualization of Embeddings using PaCMAP, with HDBSCAN Clustering
Dimension 2
Dimension 1
10
0
−10
−20
−20
−10
0
10
20
Text: processes of printing euro banknotes and minting euro coins with a national side, their
Cluster: 10_euro_currency_exchange
Text: The euro is a binding
Cluster: 10_euro_currency_exchange

LLM summaries

- All running on local LLMs for privacy reasons.

Q: Provide summaries of the most unusual document in cluster*

Response:

Scam Involving Timeshare Conversion:

- Sale of a timeshare condo as a campground membership.
- Down payment of $10,000 without delivery of promised timeshare product.
- Denial of refund despite a contract addendum.
- Accusation of credit card fraud protection guarantee non-honoring by Chase.
- Loss of significant funds without receiving promised item or refund.
- Seeking governmental and legal assistance due to fraud.

Unauthorized Charges and Identity Theft:

- Unauthorized changes to bank account details and subsequent unauthorized transactions.
- Denial of dispute by bank despite evidence of fraudulent activity.
- Credit bureaus listing unauthorized transactions as late payments and charge-offs.
- Loss of funds due to identity theft and fraudulent charges.
- Seeking assistance to recover stolen money and address unauthorized account changes.

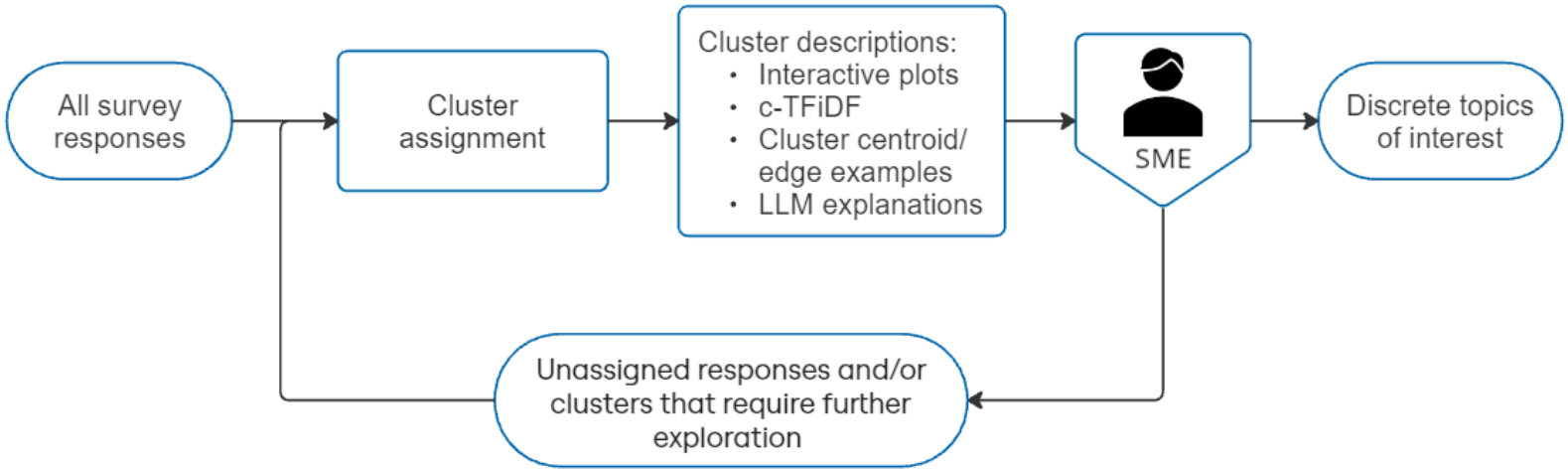
Fraudulent Transactions and Disputes:

- Merchant fraudulently disputing charges and misrepresenting to credit bureaus.
- Credit bureaus listing unauthorized transactions as legitimate.
- Credit card companies failing to resolve disputes and continuing to charge customers.
- Scenario involving a guest with a Capitol One card charged twice for a non-existent hotel stay.
- White Dial Stainless Steel watch purchase fraudulently disputed due to account access by an unauthorized third party.

* More prompt engineering behind the scenes

Iterative refinement

- Cluster assignments – no matter what algorithm you use – can be a mixed bag.
- More problematic: sometimes they're hard to explain!
- Having SMEs "in the loop" helps here.



Iterative refinement

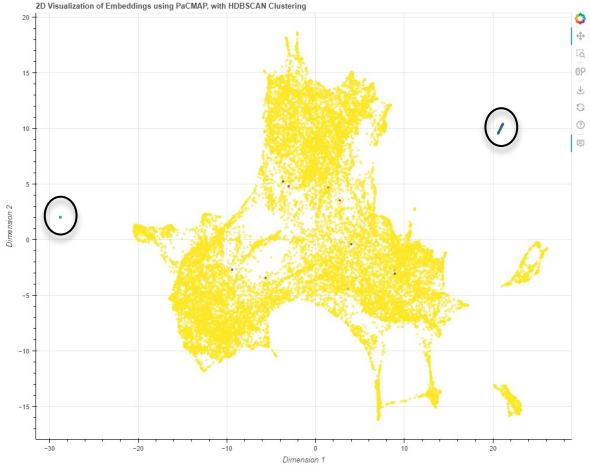
- Cache discrete clusters (circled)
- Rerun clustering on large central cluster

→

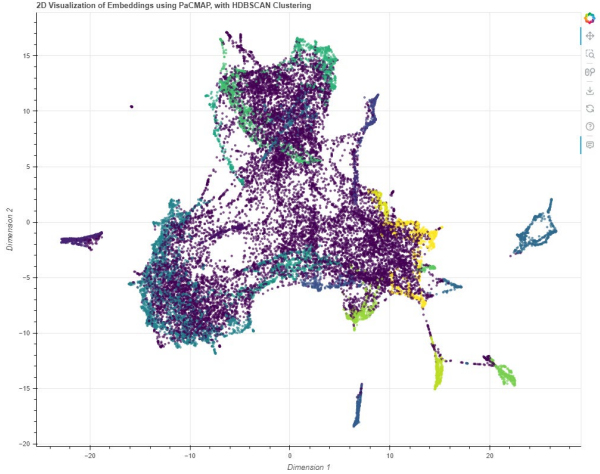
- Remove noise cluster (for downstream re-clustering)
- Focus in on clustered response

→

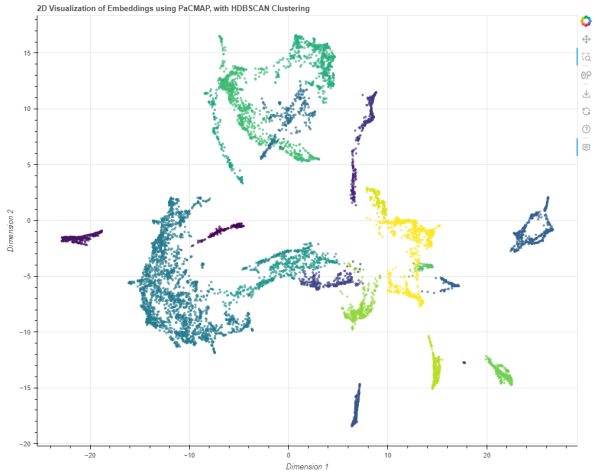
Run analysis on 20 clusters (plus 2 cached clusters from first iteration)



sentences = 23049
number of clusters = 3
noise cluster count = 12



sentences = 22882
number of clusters = 20
noise cluster count = 13818



sentences = 9064
number of clusters = 20
noise cluster count = 0

Recommendations

- Sit down with your SMEs and explain all your charts.
- Let them see the data underlying any LLM summaries – the chain of trust is important!
- Automate the process where possible – tight feedback loops are good.

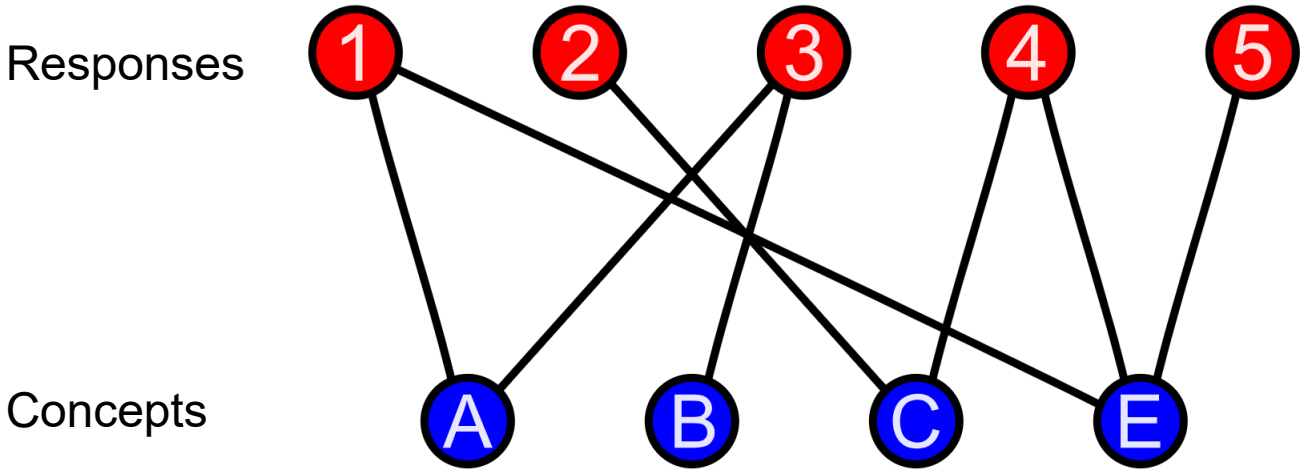
- Feed the refined clusters back into the gradual supervision pipeline.
  - Usually, iterative refinement takes place somewhere between the un- and weakly-supervised steps.

Concepts

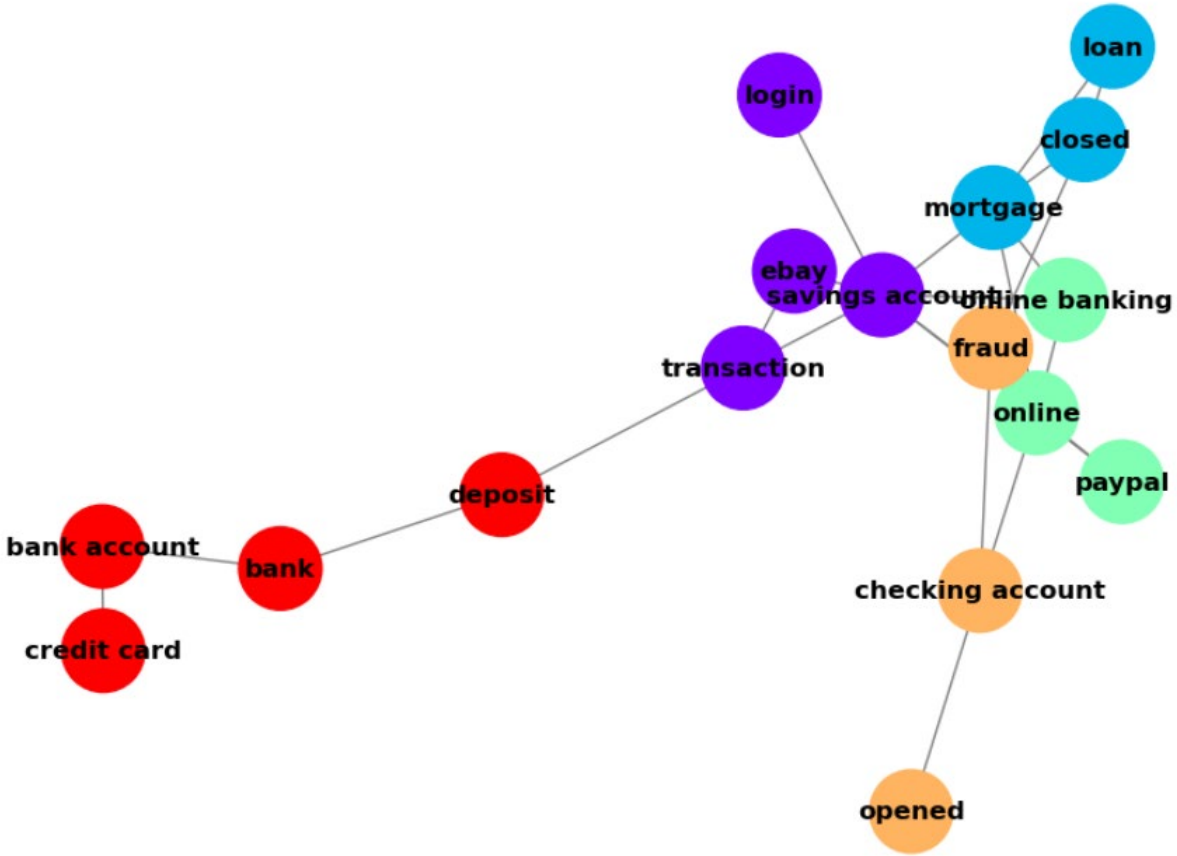
- When the output of gradual supervision is a set of classifiers, we call the classifier output *concepts*.
- Reporting on concepts is straightforward: bar charts, line graphs, etc.
- But what about more complex interactions?

Concept co-occurrence

- Each text response is associated with a set of concepts. This forms a “bipartite graph”. We can “project” these into concept-concept matrices.



Concept co-occurrence networks



Interpreting concept co-occurrence networks

- An edge means that two concepts are related (i.e. many responses are tagged with both concepts).
- Clusters on this graph can reveal more complex relationships that bar charts or group-bys wouldn't!
- If we see tight clusters (e.g. cliques) it's a sign that three separate concepts might be usefully thought of as a group.
  - In some cases!

Conclusions

- Guided supervision
  - Work closely with your SMEs
  - Use the whole gamut of techniques from un- to weakly- to fully-supervised.
- Cluster exploration
  - Important for reducing potential bias from basic topic summarization.
- Refinement
  - Build tight interactive loops with your SMEs to allow them to refine group assignments.
  - Use LLMs to reduce cognitive load on SMEs, but ensure they have access to the underlying data.
- Interactions
  - It’s worth plotting a co-occurrence network if you’ve got a lot of concepts that might be interrelated.