Enhancing Big Data Classification with Zero-Shot Learning Models

A Large Classification Problem

Francis Smart - Censeo Consulting, Michigan State University PhD (August 2024)
Issa Abboud - General Services Administration

Contractors Participating in TDR Must Submit Transactional Data

If the contractor participates in TDR they must electronically report transactional data (GSAR clause 552.216-75)

The transactional data consists of 12* mandatory data fields

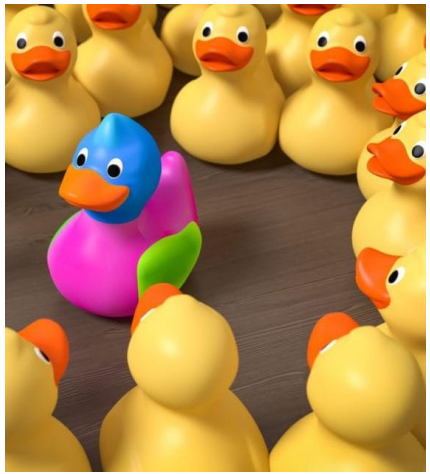
the CSP disclosures and the PRC will no longer be required

This reporting is intended to enhance the government's ability to conduct data analysis to make smarter purchasing decisions through the sharing of information.

* The SIN field was added after the GSAR

TDR Can Empower the AWF

Why Transactional Data Reporting (TDR)?



TDR is Unique:

- Unequaled level of granularity
- Enhances data analysis capabilities
  - Historical pricing trends
  - Seasonal patterns
- Empower the acquisition workforce
  - Leverage to better understand and collaborate
  - Make informed procurement decisions

TDR is tightly aligned with priority 3 (Managing the Business of Government) in the President's Management Agenda (PMA)

Low Quality Inhibits Our Ability to Buy Smarter

Observations:

- GSA uses SINs (Special Item Numbers) to categorize products or services. Contractors report sales under a particular SIN.
- GSA has three types of SINs: Some are strictly products, others are strictly services, others are both products and services.

Problem Statement:

SINs that span both products and services make it challenging to classify a reported item as a product or a service. This in turn diminishes our ability to analyze reported transactions, and inhibits us from buying smarter and more efficiently.

Item Classification Increases Data Quality

Objective: Design a classification model to identify reported items as either a product or a service

Successful classification of TDR items:

- Increases the overall quality of the dataset
- Enables us to better analyze products and services sold to the government and make informed procurement decisions

Need to Classify Big Data

Transactional Reporting Data: Since January 2022

- 9 million + rows
- $34.9 billion in reported transactions
- 1 million unique descriptions since 2022

How do we make sense of this kind of so much data

- 835 SIN Numbers (Special Item Number) allow for classification into general categories.

Top 10 Categories in Terms of Value

Level 1 Category	Level 2 Category	Total Value	Total Value%	Total Count	Total Count%	Unique Desc.
IT	Products	$12,125,978,028	34.7	688,723	7.4	153,126
IT	IT Professional Services	$9,480,300,635	27.1	623,338	6.7	56,922
Professional Services	Technical Engineering Services	$3,118,017,328	8.9	146,659	1.6	5,342
Professional Services	Financial Services	$1,941,048,294	5.6	37,691	0.4	3,345
Facilities Construction	Facility Related Services	$1,652,209,951	4.7	176,736	1.9	26,365
Industrial Products Services	Hardware Tools	$1,568,513,519	4.5	3,872,690	41.8	530,486
Professional Services	Management Advisory Services	$1,191,534,147	3.4	66,915	0.7	3,354
Human Capital	Talent Development	$932,882,627	2.7	161,464	1.7	2,421
Office Management	Office Management Products	$657,804,017	1.9	2,810,121	30.3	186,082
Industrial Products Services	Industrial Products Install / Maintenance / Repair	$523,301,399	1.5	26,780	0.3	2,763
Transportation and Logistics Services	Logistics Support Services	$419,608,017	1.2	11,633	0.1	614

But SIN number does not always match records of interest or is sufficiently specific.

Validations by Services or Product

Data validations are quite different between products and services:

- Products validate: unit of issue, price per unit, manufacturer name, manufacturer part number against GSA Advantage Catalog, check if price per unit is competitive
- Services: labor code (if available), price per hour, description for more details

	Total Value	Total Count	Unique Desc.	Total Value%	Total Count%
Product	$14,090,063,030	5,955,278	843,078	40.3	64.2
Services	$12,833,975,892	836,204	48,130	36.7	9
Product or Services	$7,530,701,813	2,350,251	221,180	21.6	25.4
Unknown	$487,190,861	128,098	22,529	1.4	1.4

Validations by Services or Product

	Total Value	Total Count	Unique Desc.	Total Value%	Total Count%
Product	$14,090,063,030	5,955,278	843,078	40.3	64.2
Services	$12,833,975,892	836,204	48,130	36.7	9
Product or Services	$7,530,701,813	2,350,251	221,180	21.6	25.4
Unknown	$487,190,861	128,098	22,529	1.4	1.4

Vendors Match the Wrong SIN:

- Some rows are given a Service SIN but are products
- Some rows are given a Product SIN but are Services

Predicting Product or Service

So how do we handle this?

- The first take would make the data look like a good fit for standard NLP model using “Service” and “Product” categories as correctly classified and predict the “Service or Product” category rows from them.
- Unfortunately, initial experimentation indicated low performance likely due to large amounts of heterogeneity in descriptions and widespread SIN misuse.

At this point we decided to experiment with a pretrained Zero-Shot Classifier.

Zero-Shot Classification

- Zero-shot learning (ZSL) is a classification task where none of the classes seen during training are present at test time. It relies on semantic understanding to infer the characteristics of unseen classes.

- Application: Useful for scenarios where it is impractical to have labeled data for every category.

- Importance: Enables models to generalize to new tasks without needing explicit examples.

Introduction to BART

- BART (Bidirectional and Auto-Regressive Transformers) combines the benefits of both autoencoder and autoregressive models.

- Architecture: Built using the Transformer model, consisting of a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT).

- Usage: Effective for a range of NLP tasks including text generation, comprehension, and translation.

Introduction to bart-large-mnli

- Base Model: Based on BART model architecture.

- Fine-Tuning: Specifically fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus.

- Purpose: Designed to predict textual entailment (i.e., whether a given text implies a hypothesis), which is critical for zero-shot classification.

Zero-Shot Classification with `bart-large-mnli`

- Mechanism: Utilizes textual entailment to predict the likelihood of hypotheses (labels) given a text, without previously seeing examples of those labels.

  - Application: Can dynamically handle any label provided at runtime, making it extremely flexible across different domains.

  - Strength: Leverages deep semantic understanding, making it robust to variations in text input.

Applications and Benefits

- Versatility: Suitable for content categorization, intent detection, and more without specific training data.

- Efficiency: Reduces the need for extensive labeled data sets and additional model training.

- Scalability: Easily adaptable to new and emerging categories or languages.

But first how do we make the problem more tractable?

1. Group rows by description.
2. Select only the Top 10,000 rows (sort by Value or Count)

Sorted by Total Value

Row Bin	Total Value	Total Value Cumulative	Total Value % Cumulative	Total Count	Total Count Cumulative	Total Count % Cumulative
Less than 10^3	$18,389,802,053	$18,389,802,053	53	340,874	340,874	4
10^3 to 10^4	$9,862,160,653	$28,251,962,706	81	894,089	1,234,963	13
10^4 to 10^5	$5,819,797,264	$34,071,759,970	98	4,026,156	5,261,119	57
Greater than 10^5	$870,186,862	$34,941,946,832	100	4,008,712	9,269,831	100

Sorted by Row Count

Row Bin	Total Value	Total Value Cumulative	Total Value % Cumulative	Total Count	Total Count Cumulative	Total Count % Cumulative
Less than 10^3	$3,357,511,874	$3,357,511,874	10	2,176,527	2,176,527	23
10^3 to 10^4	$5,434,792,397	$8,792,304,271	25	2,532,571	4,709,098	51
10^4 to 10^5	$15,218,365,787	$24,010,670,058	69	2,661,169	7,370,267	80
Greater than 10^5	$10,931,276,774	$34,941,946,832	100	1,899,564	9,269,831	100

When looking to Classify Text We Need Categories

There is ambiguity in the naming of some important features.

- Products (Tangible Consumables)
- Professional Services (Labor Hours)
- Firm Fixed Price

What about?

- Cellular service
- Warranties
- Installation Costs of Products
- Etc.

Our Categories (After some exploration)

Description	Category	Count	Out of 10^4
fees such as delivery fee, or expedited fee	Other	236	2.36%
non-labor hour service such as cell phone, internet,	Product	1,358	13.58%
products	Product	2,879	28.79%
professional human service	Labor Hour	3,190	31.90%
software	Product	2,337	23.37%

Results

Threshold Weight	Count	Overall Accuracy %	Labor Hour Accuracy %	Labor Hour Identified %
0	412	69.17	82.52	82.52
0.1	412	69.17	82.52	82.52
0.2	412	69.17	82.52	82.52
0.3	409	69.68	82.93	82.52
0.4	405	70.37	82.93	82.52
0.5	301	76.74	86.71	66.5
0.6	215	81.86	90.27	49.51
0.7	144	83.33	89.33	32.52
0.8	59	83.05	80.77	10.19
0.9	17	88.24	75	1.46
0.95	6	100		0

- Hand coded top 500
- 412 were discernible categories
- 170 were labor hour
- If we select threshold based on max accuracy we lose 33 percent identification

Next Steps

- Fine Tune
- Scale up (AWS - SparkNLP)
- Move from 80% of value to 80% of counts.
- Set up a flow to identifying various other goods or services of high interest which have high values and inconsistent part numbers.
  - Desktop computers
  - Laptops
  - Software suites

Thank you for your time!

Francis Smart
Censeo Consulting
FSmart@CenseoConsulting.com
Francis.Smart@gsa.gov

Issa Abboud
General Services Administration
Issa.Abboud@gsa.gov