



**STATISTICS**



# **Coding Time Series With Machine Learning**

**3RD IRVING FISHER COMMITTEE WORKSHOP  
ON “DATA SCIENCE IN CENTRAL BANKING”**

**BANCA D’ITALIA - OCTOBER 18, 2023**

Ayoub Mharzi – Data Scientist  
IMF Statistics Department

# Context

Increasingly available  
text data

Surveys



Data tables



Websites



Challenges and  
opportunities

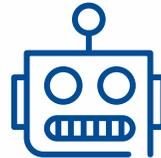
Nonstandard  
terminology



Outdated  
processes



New technologies



Objectives

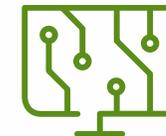
Understand  
the data



Explore new  
approaches



Design solutions



# What is “coding time series”?

- IMF member countries publish economic time series data in their National Summary Data Pages (NDSPs)
- IMF staff map these data to the internal Catalogue of Time Series (CTS) to ingest in IMF database.
- This is referred to as “coding time series.”

DATASTRUCTURE	IMF:ECOFIN_DSD(1.0)	Datastructure		
DATASTRUCTURE_NAME	ECOFIN Data Structure Definition	Datastructure Name		
DATA_DOMAIN	CGO	Dataset		
REF_AREA	KW	Country		
COUNTERPART_AREA	_Z	Counterpart area		
UNIT_MULT	6	Scale = Million		
FREQ	A	Frequency = Annual		
COMMENT	Source: <a href="https://www.cbk.gov.kw/en/stati">https://www.cbk.gov.kw/en/stati</a> Source / Observation status			
Descriptor_Alt	Descriptor	INDICATOR	2009	2010
	General public services and defence			
	Public order and safety			
	Education affairs and services			
	Health affairs and services			
	Social security and welfare affairs and s			
	Housing and community affairs and serv			
	Recreational, cultural and religious affa			
	Fuel and energy affairs and services			
	Agricultural affairs and services			
	Manufacturing affairs and services			
	Transportation and communication affa			
	Other economic affairs and services			
	Expenditures, not classified by function			

# Challenges



Time consuming and manual.



**Finding similarities is not simple**  
when new wording is used for time series descriptors.

**Not always straightforward.**  
E.g., “Health, affairs and services”:  
218 indicators with the word “health”;  
32 with the word “service.”

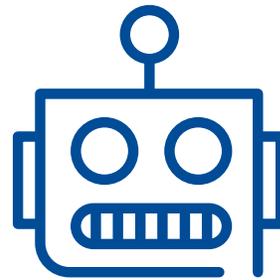
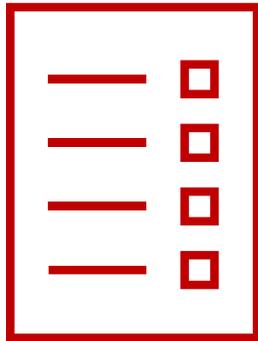


**Subject to Human Error**

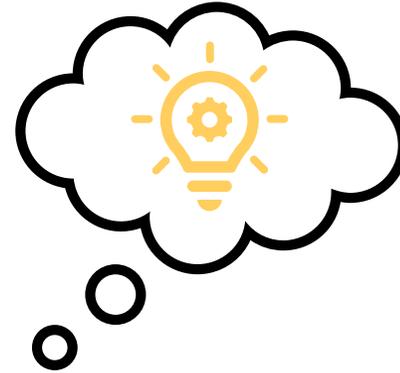


# Hypothesis

Country upload files with missing codes



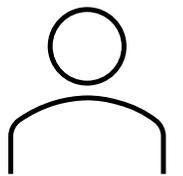
ML Model



Files with codes predicted by ML



Feedback



User

# Data processing



~142k coded indicators  
from 53 countries across  
4 macroeconomic sectors



Build a master dataset  
based on already coded  
data files to train/test



Light data processing and  
cleaning approach to  
allow reproducibility and  
scalability



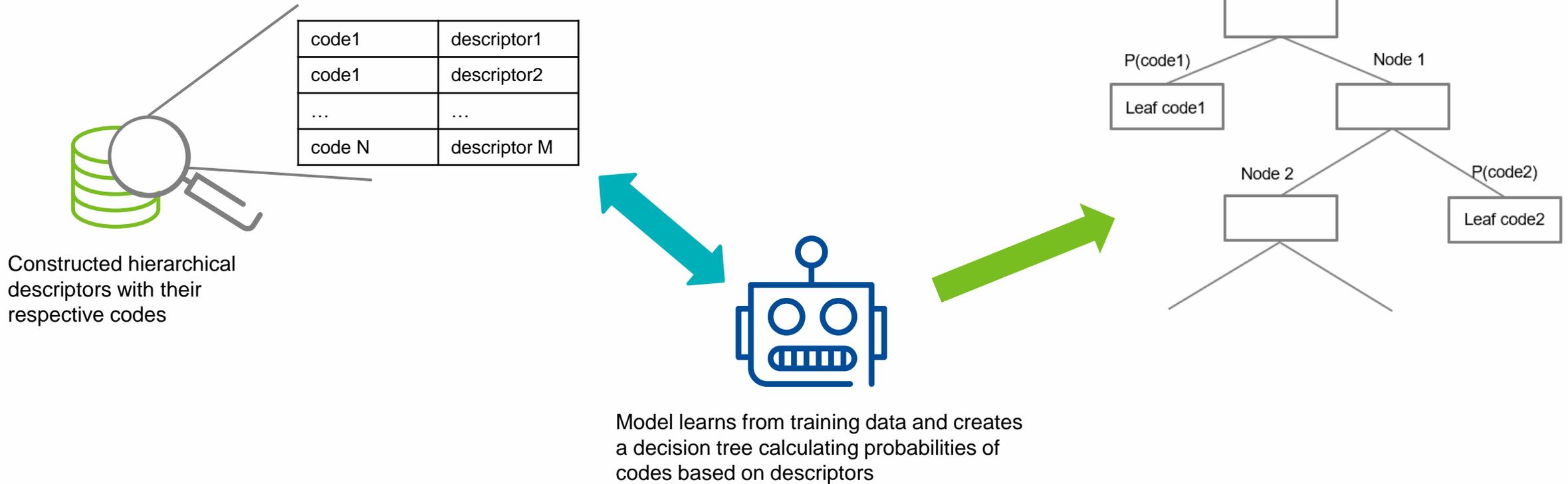
A few data files were in  
French, Spanish and  
Portuguese

- **Remove country-specific codes.**
- **Construct descriptors hierarchy.**
- **Keep only alphanumeric characters and remove stop words** (“a”, “the” etc.)
- **Keep available metadata fields** such as data domain, frequency, etc.
- **Train and Test data:** 90/10 split for training and testing/validation

# Methodology

- Our data have specific terminology, a **supervised ML approach** is used.
- Start with proof of concept (PoC) to test the **feasibility and adequacy** of several ML models and feature extraction techniques (TF-IDF, word embeddings, Skip-Gram, logistic regression, etc.)
- Use an open-source text classification library, **fasttext**
- Text classification with Skip-Gram model and 3 n-gram
- Vectors dimension: 100, Epochs: 20

# High-level underlying logic

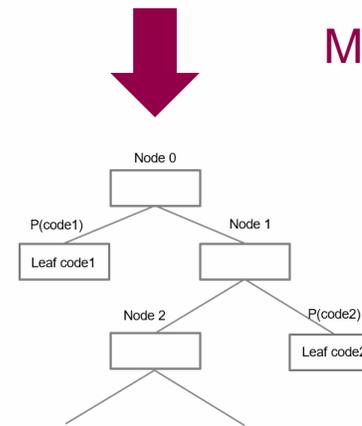


# Overall process

## 1 Data Processing



## Model 2



## 3 Rank codes based on their probability

Code N	Probability N
code2	Probability 2
...	...
code 1	Probability 1

## 4 Returning top 5 codes with highest probability

<b>GEL_G14_CG</b>	Government and Public Sector Finance, Expenditure by COFOG, Central Government, Health [2014 Manual]
...	...
...	...
...	...
...	...

# Results

- Precision ranging by data domain, majority of domains precision around 90%
- Recall ranging by data domain, majority of domains recall around 90%
- Cases of low precision and recall due to:
  - ◆ Low number of time series for certain domain in training dataset
  - ◆ Original data file's structure not allowing to construct full hierarchy indicators
- For a data file with ~500 time series, the model takes about 2 seconds to return predicted codes

Domain	N	precision	recall
balance payments bpm	5,014.0	0.965	0.966
m&b central bank	1,309.0	0.915	0.914
m&b depository corporations	1,293.0	0.92	0.922
international investment position bpm	1,205.0	0.975	0.973
direction trade statistics	771.0	0.966	NA
financial soundness indicators sectoral financial statements	544.0	0.965	0.963
financial soundness indicators institution deposit takers ofcs	533.0	0.32	0.319
national accounts gross domestic product	398.0	0.885	0.872
financial soundness indicators data report form	348.0	0.965	0.963
m&b financial corporations	344.0	0.925	0.924
balance payments	337.0	0.99	0.991
financial access survey	267.0	0.86	0.861
international investment position	247.0	0.98	0.980
merchandise trade	246.0	0.69	0.679
consumer price index	224.0	0.915	0.915

# Achievements

- Use open-source tools to build a solution
- Built a specific vocabulary and train the model
- Light data processing allowing reproducibility and scalability of the tool
- Achieved for most domains a high-level Precision and Recall
- Create an **R package** with the documented code available for sharing upon request
- Build an **R-Shiny app** for end users based on received feedback

# Shiny App

The screenshot displays the CTS Coding Tool interface, which is a Shiny App. The interface is organized into three overlapping panels, each showing a different view of the application. The top panel shows the 'Export Options' section, which includes a sidebar with navigation links (Introduction, Upload, CTS Classification, Export Results, CTS Reference) and a main content area with the following settings:

- Export Options:**
  - Select additional columns to export:  CTS\_DESCRIPTOR,  DATA\_DOMAIN\_NAME,  METHODOLOGY
  - Select export format:  CSV,  Excel
  - CTS codes below threshold:  Add,  Don't add
  - Export button

The bottom panel shows the 'Export Data Preview' section, which includes a search bar and a table of results. The table has the following columns: id, DESCRIPTOR, predicted\_code, CTS\_DESCRIPTOR, DATA\_DOMAIN\_NAME, and probability. The data is as follows:

id	DESCRIPTOR	predicted_code	CTS_DESCRIPTOR	DATA_DOMAIN_NAME	probability
1	Final Consumption Expenditure [FCE]	NCGG	National Accounts, Expenditure, Gross Domestic Product, Final Consumption Expenditure, Public Sector (General Government and Public Corporations), General Government, Nominal	NAG	0.963
2	General Government	BEFPDGCAL_BP6	Balance of Payments, Memorandum Items, Exceptional financing, Portfolio investment, liabilities, Debt securities, General government, Cancellation of arrears, Interest/coupon [BPM6]	NAG	0.838
5	Gross Fixed Capital Formation	NFI	National Accounts, Expenditure, Gross Domestic Product, Gross Capital Formation, Gross Fixed Capital Formation, Nominal	NAG	0.995
8	Gross National Expenditure	NGDE	National Accounts, Expenditure, Memorandum Items, Gross Domestic Expenditure, Nominal	NAG	0.872
10	Imports	TMG	External Trade, Imports, Goods, Value	NAG	0.888
11	Statistical Discrepancy	NSDGBP	National Accounts, Expenditure, Memorandum Items, Statistical Discrepancy in Gross Domestic Product, Nominal	NAG	0.891

At the bottom of the preview, it shows 'Showing 1 to 6 of 6 entries' and navigation buttons for 'Previous', '1', and 'Next'.

# Next steps

- Incorporate user feedback into the end-to-end pipeline
- Officially roll out the tool into production and implement monitoring mechanisms
- Test the model and solution for out-of-scope time series
- Make the R package sharable upon request

# Lessons learned

- **NLP/ML is a fast-changing field:** methods such as TF-IDF or Random Forest, although still used as baseline, are increasingly challenged by more recent methods such as word embedding, transformers, and other deep learning models.
- **Generative AI is changing the landscape of NLP:** considering if it is worth developing in-house solution or opt for off-the-shelf solutions such as ChatGPT.
- **Importance of the maintenance strategy:** Preparing in advance the monitoring and re-training plan is crucial for the sustainability of the model and can increase the acceptance and facilitate the transition from PoC to production.
- **Machine Learning solutions take time to implement to reach robust and high-quality results.**
- **Collaboration and knowledge sharing can have a big impact :** peer organizations can face similar challenges and we believe a solution like ours could be adapted and implemented by others.

# Useful References



- Text and classification
  - Model retraining



Measure (2017): use artificial intelligence (AI) to map worker injury to internal coding



Statistics  
Canada

Evans et al (2021), "Need for Speed: Using fasttext (Machine Learning) to Code the Labor Force Survey"

# Thank You

Team: Alberto Sanchez, Alessandra Sozzi, Ayoub Mharzi, Lamyia Kejji and Yamil Vargas

Ayoub Mharzi: [Amharzi@imf.org](mailto:Amharzi@imf.org)

Analytics Team: [STA-Analytics@imf.org](mailto:STA-Analytics@imf.org)