Leveraging Open-Source Language Models for Automatic Codification of Employment and Economic Activity in Household Surveys

Alejandro Pimentel

12/Dec/2024

Introduction

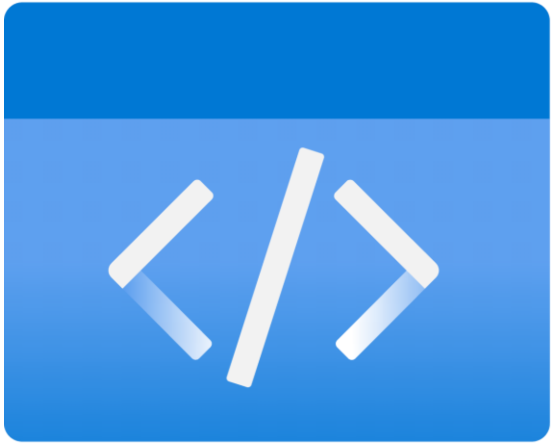
Enhancing Efficiency in Statistical Processes

Focus: Occupation (SINCO) and Economic Activity (SCIAN) variables

Codification Framework

NAICS: North American Industry Classification System

SINCO: National Occupational Classification System



Surveys

ENIGH: National Survey of Household Income and Expenditure

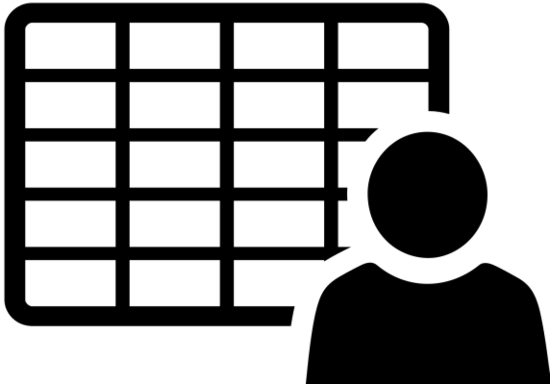
- Quarterly
- 40000 records
- 260 coders
- Decentralized

ENOE: National Survey of Occupation and Employment

- Biennial
- 30000 records
- 10 coders
- Centralized

EIC: Intercensal Survey

- Quinquennial
- More than 1 million records
- 600 coders
- Centralized



The Challenge

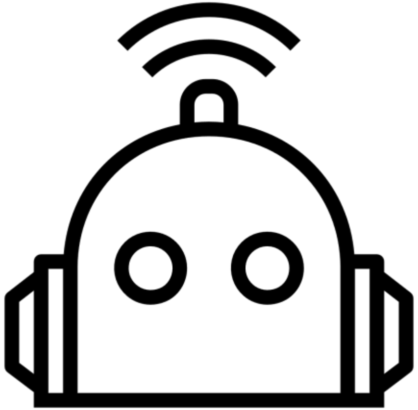
Manual Coding Limitations
Labor-intensive and time-consuming

Requires extensive training and resources

Prone to human error

The Solution

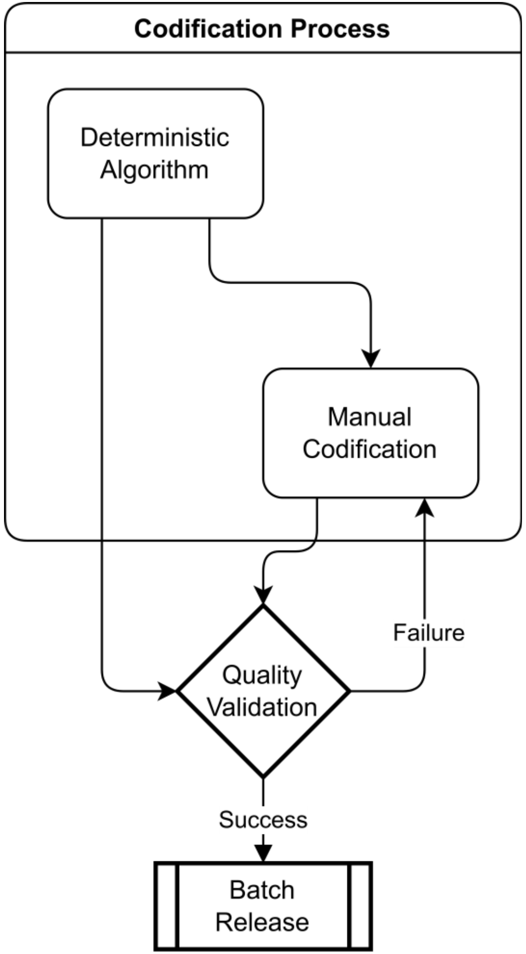
Using AI to Automate Coding Tasks
Implementation of AI algorithms
Reduction in manual workload
Improvement in accuracy and consistency



Original Methodology

Process Development and Evaluation Phases:

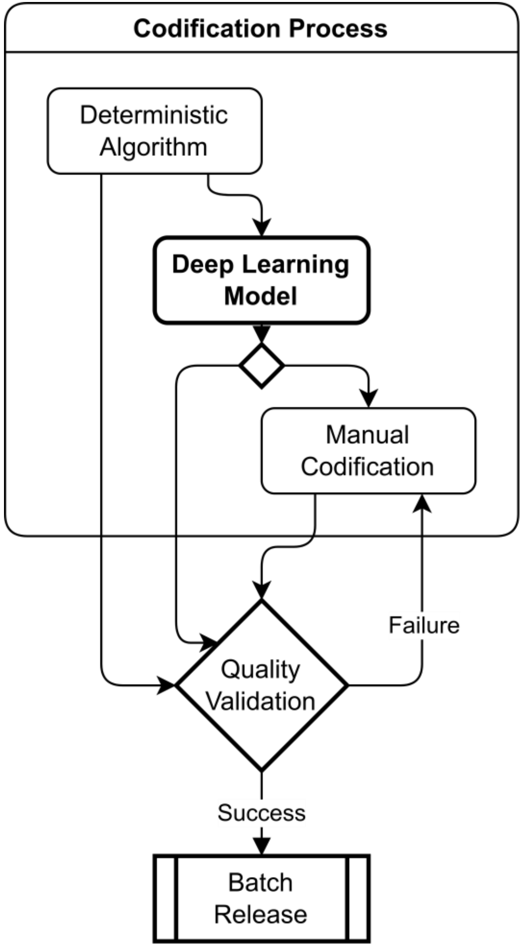
1. Deterministic algorithms
2. Manual Coding
3. Quality validation



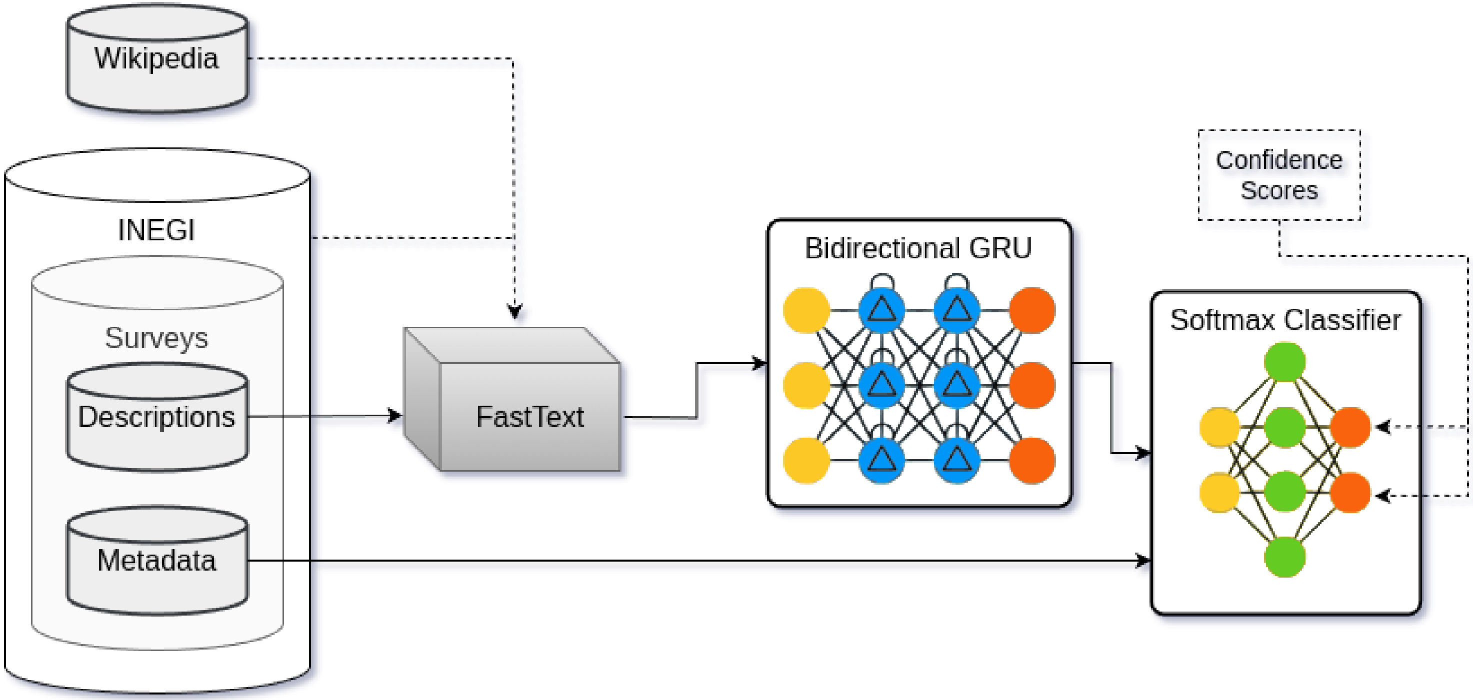
Methodology

Process Development and Evaluation Phases:

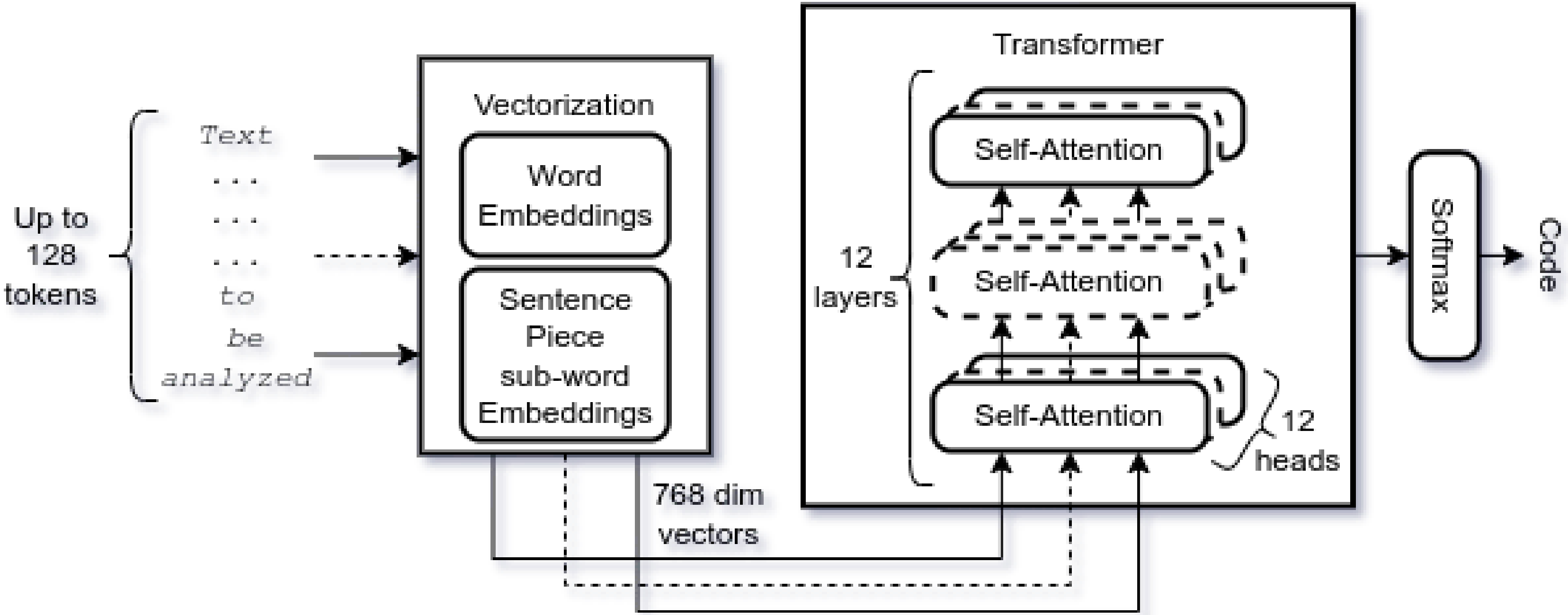
1. Deterministic algorithms
2. **Artificial Intelligence model**
3. Manual Coding
4. Quality validation



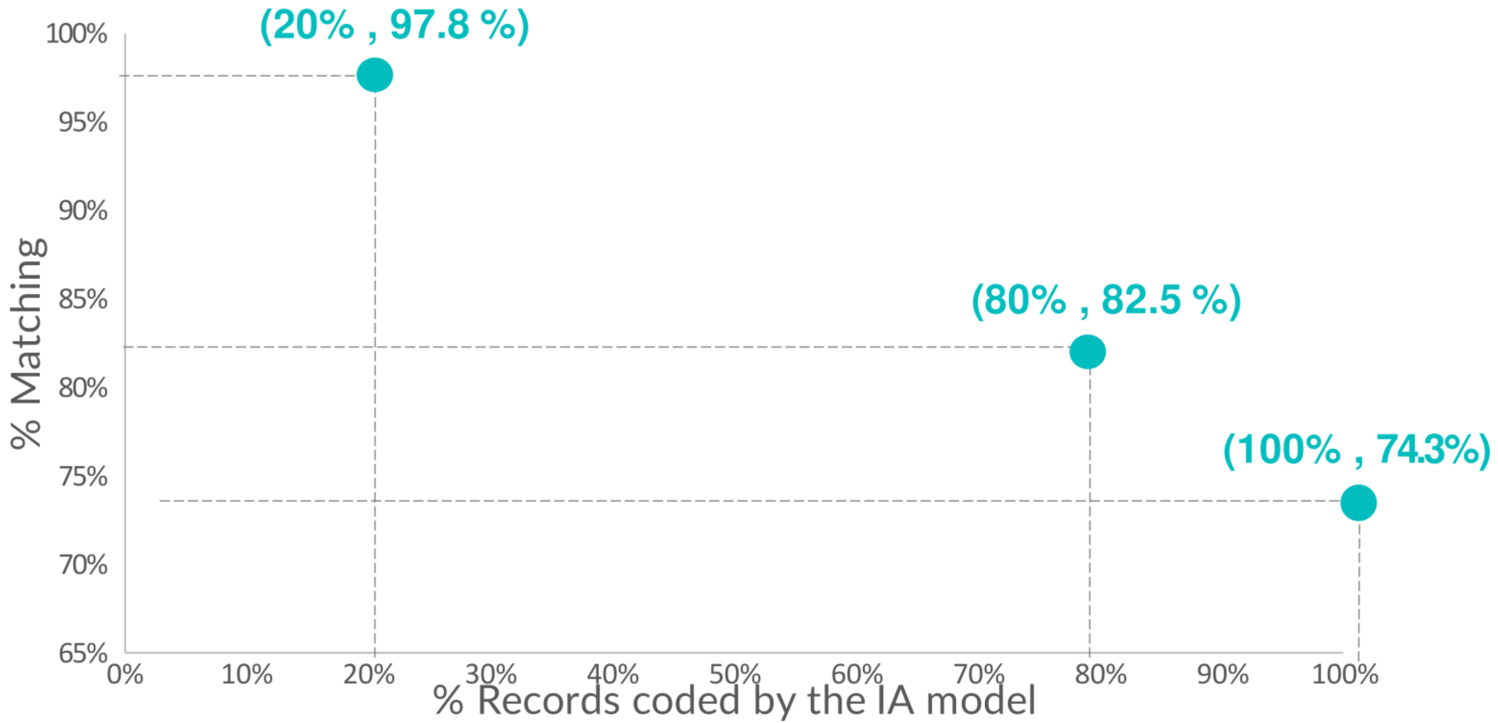
ENIGH Survey



ENOE Survey


Up to
128
tokens
Text
. . .
. . .
. . .
to
be
analyzed
Vectorization
Word
Embeddings
Sentence
Piece
sub-word
Embeddings
768 dim
vectors
Transformer
12
layers
Self-Attention
Self-Attention
Self-Attention
12
heads
Softmax
Code

Threshold Trade-off



INEGI

ENIGH Threshold Analysis - Occupation

Threshold	Val Savings	Val Matching	Matching	Quality	Savings	*h*	p-value
0.472565	90.0%	70.7%	73.2%	71.4%	89.3%	0	1
0.586216	80.0%	74.5%	77.4%	75.2%	79.5%	0	0.9752
0.705673	70.0%	78.4%	81.4%	79.3%	69.4%	0	0.0810
0.808395	60.0%	81.9%	85.6%	83.2%	59.0%	1	6.0133e-04
0.889082	**50%**	**85.4%**	**89.4%**	**86.5%**	**48.2%**	1	6.8188e-05
0.943088	40.0%	88.5%	92.6%	89.3%	37.9%	1	7.6920e-05
0.973509	30.0%	91.4%	95.3%	91.8%	28.0%	1	0.0023
0.989804	20.0%	94.6%	97.5%	93.7%	19.0%	0	0.0662
0.997738	10.0%	96.8%	98.9%	95.6%	9.2%	0	0.5000

ENIGH Threshold Analysis - Activity

Threshold	Val Savings	Val Matching	Matching	Quality	Savings	*h*	p-value
0.534222	90.00%	76.70%	77.1%	76.8%	89.7%	0	0.9850
0.677070	80.00%	80.40%	81.1%	80.4%	79.6%	0	0.6535
0.793442	70.00%	83.90%	84.9%	83.9%	70.1%	0	0.0797
0.881765	60.00%	87.00%	88.8%	87.2%	60.2%	1	0.0054
0.936173	**50%**	**89.7%**	**92%**	**89.9%**	**50.5%**	1	0.0034
0.970834	40.00%	92.50%	95.1%	92.7%	40.4%	1	0.0173
0.988363	30.00%	94.80%	97.7%	95.0%	28.9%	0	0.1279
0.996108	20.00%	96.90%	98.9%	96.5%	18.1%	0	0.1334
0.999112	10.00%	97.90%	99.8%	97.8%	8.3%	0	0.2500

INEGI

ENOE - ENIGH Comparison

Application on full data.

	Activity	Occupation
ENIGH	76.7	70.7
ENOE	84.8	77.4

ENOE - ENIGH Comparison

Application on half the data.

	Activity	Occupation
ENIGH	89.7	85.4
ENOE	98.4	94.3

Final Remarks

- Significant improvements in coding accuracy and efficiency using advanced models like BERT and FastText.
- Our models have reached or exceeded the quality levels of manual coding, with notable improvements in the ENOE survey.
- Fine-tuning thresholds has shown promising results in balancing quality and savings.
- The success of the models is strongly linked to the quality and curation of the training data.

Next Steps

- Advancing with LLMs
- Creation of a comprehensive ground truth database


THANKS
Conociendo México
800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx
INEGI Informa