Natural Language Processing for the variables of Occupation and Economic Activity

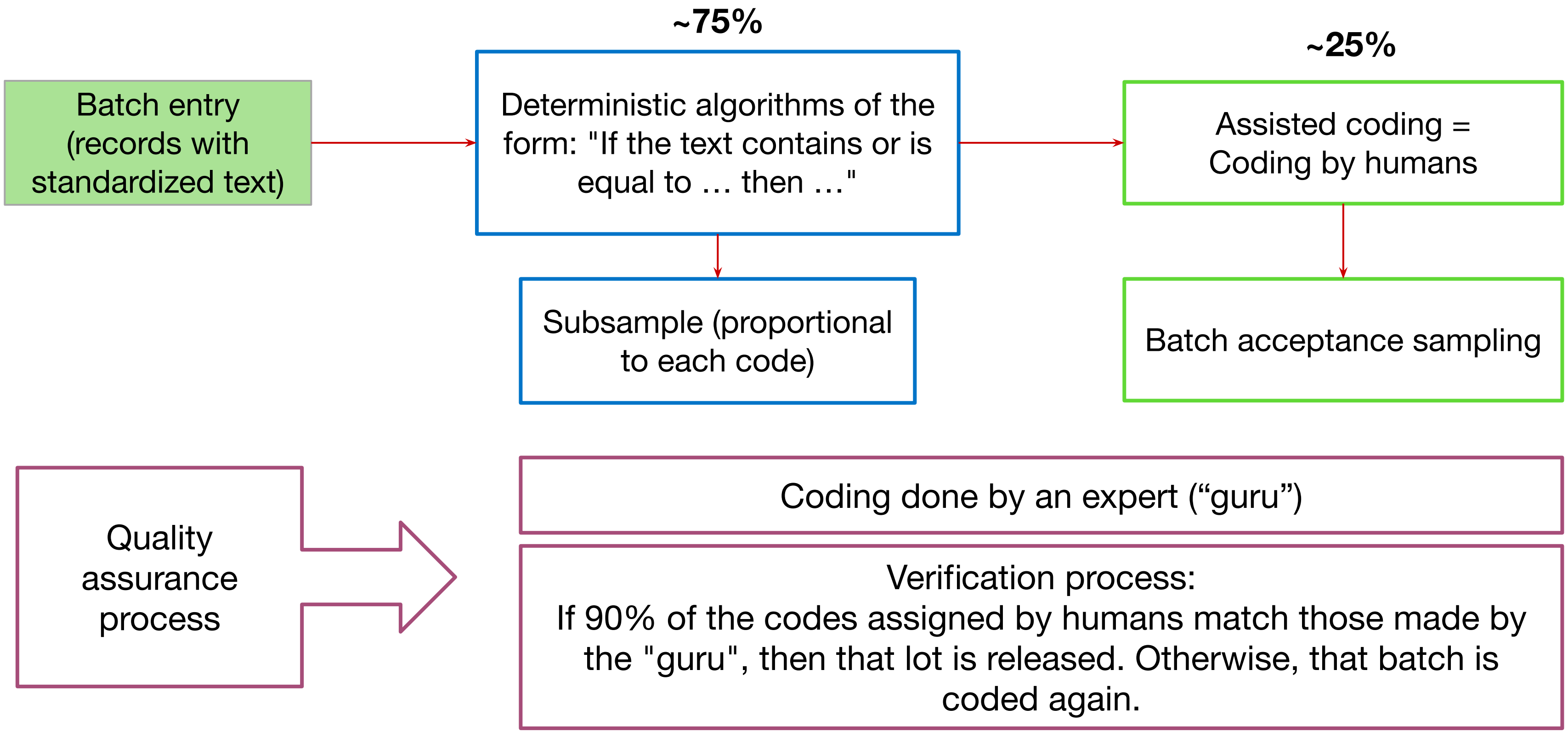
José Alejandro Ruíz Sánchez / Jael Pérez Sánchez

Objectives

○ Evaluate the incorporation of Natural Language Processing (NLP) techniques within the productive coding flow:

- Reduce the workload of human coders
- Reduce coding time
- Maintain or improve encoding quality

Current process



~75%

~25%

Batch entry (records with standardized text)

Deterministic algorithms of the form: "If the text contains or is equal to … then …"

Subsample (proportional to each code)

Assisted coding = Coding by humans

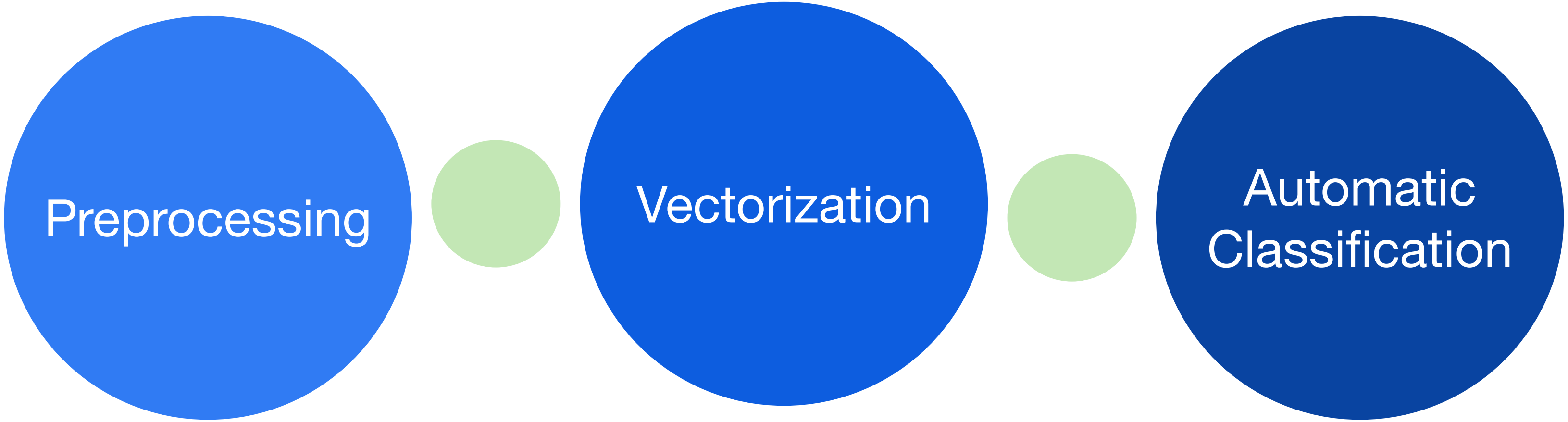
Batch acceptance sampling

Quality assurance process

Coding done by an expert (“guru”)

Verification process:
If 90% of the codes assigned by humans match those made by the "guru", then that lot is released. Otherwise, that batch is coded again.

Stages of the NLP Text Classification Process



Information used

We use the 2018 National Household Income and Expenditure Survey (ENIGH), which has 158,000 coded records.

Auxiliary text (covariates, attributes):

{Occupation: ‘MEXICAN APPETIZER COOK’}
{Tasks: ‘PREPARATION FOR SALE OF MEXICAN ANTOJITO IN LOCAL’}
{Company activity: ‘PREPARATION OF SALE OF MEXICAN ANTOJITO IN PREMISES TO THE GENERAL PUBLIC’}
{Name of the company: ‘HUARACHE MIMI’}
11 additional variables: academic level, company size, ...

Classification variable (target variable): {7221}

Hierarchical code (first 2 digits represent the sector)

Preprocessing

- Spelling corrections
- Lemmatization
- Word truncating (Stemming)
- N-grams

Treatment similar to
that
currently performed

Vectorization

Traditional

Bag of Words: count the number of tokens within each text

Distributional Embeddings: based on the co-occurrence of tokens within a window. Try to give context to the words

TF iDF: word weighing, weighs each token according to its frequency within a text and the frequency of the token within the different texts

Neural Embeddings

w2v, Fasttext: through neural networks, uses the text sequence to generate numerical vectors

ELMO, BERT: Complex Neural Network Architectures (CNN, LSTM, Mechanisms of Attention)

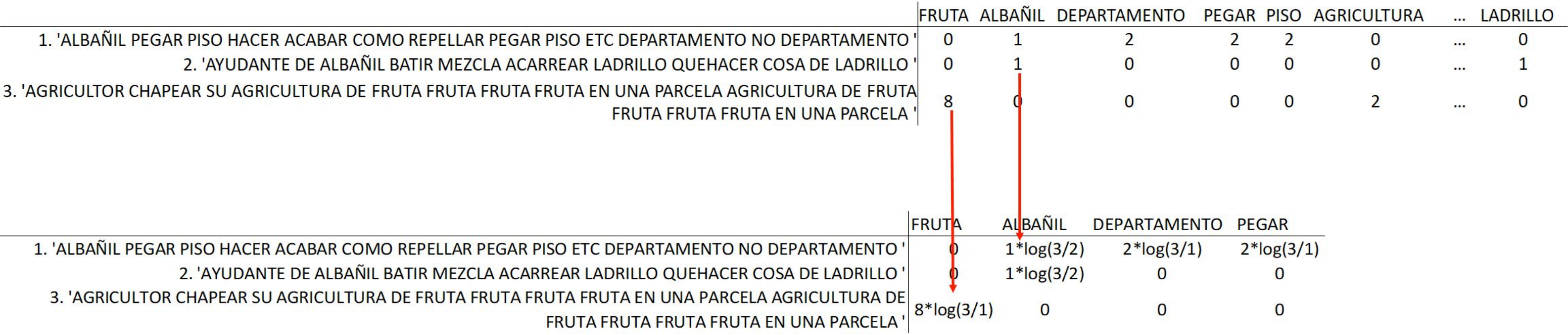
Word weighing

$$w_{i,j} = tf_{i,j} * log\left({}^{N}/_{df_i}\right)$$

$$tf_{i,j} = number\ of\ times\ the\ term\ i\ appears\ in\ document\ j$$

$$df_i = number\ of\ documents\ containing\ the\ term\ i$$

$$N = number\ of\ documents$$



	FRUTA	ALBAÑIL	DEPARTAMENTO	PEGAR	PISO	AGRICULTURA	...	LADRILLO
1. 'ALBAÑIL PEGAR PISO HACER ACABAR COMO REPELLAR PEGAR PISO ETC DEPARTAMENTO NO DEPARTAMENTO '	0	1	2	2	2	0	...	0
2. 'AYUDANTE DE ALBAÑIL BATIR MEZCLA ACARREAR LADRILLO QUEHACER COSA DE LADRILLO '	0	1	0	0	0	0	...	1
3. 'AGRICULTOR CHAPEAR SU AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA '	8	0	0	0	0	2	...	0

	FRUTA	ALBAÑIL	DEPARTAMENTO	PEGAR
1. 'ALBAÑIL PEGAR PISO HACER ACABAR COMO REPELLAR PEGAR PISO ETC DEPARTAMENTO NO DEPARTAMENTO '	0	1*log(3/2)	2*log(3/1)	2*log(3/1)
2. 'AYUDANTE DE ALBAÑIL BATIR MEZCLA ACARREAR LADRILLO QUEHACER COSA DE LADRILLO '	0	1*log(3/2)	0	0
3. 'AGRICULTOR CHAPEAR SU AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA AGRICULTURA DE FRUTA FRUTA FRUTA FRUTA EN UNA PARCELA '	8*log(3/1)	0	0	0

TF iDF

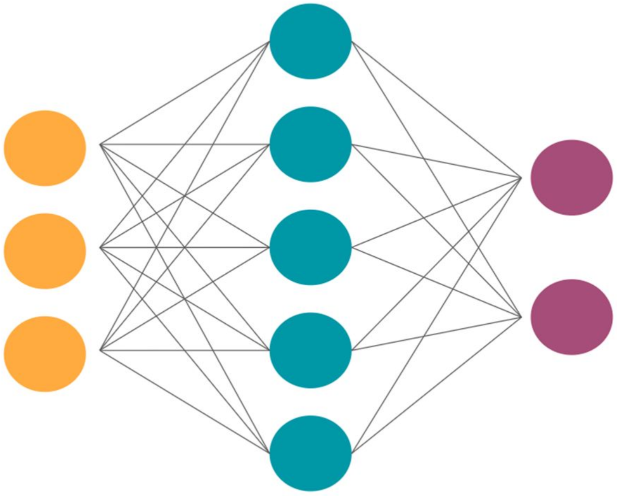
- Tends to generate sparse matrix
- For the weighing of words it doesn’t matter their order or place within the text
- Doesn’t incorporate information about the context of the word
- The number of columns is a parameter to be determined
- High computational cost with large matrices
- Stopwords tend to small values or zero
- If the number of words is restricted, it is likely that infrequent words will not be considered

w2v & fasttext: adding context

“ELABORADOR DE DULCE TIPICO DE LA REGION “

$$\underbrace{\text{ELABORADOR DE}}_{\text{context}}\ \text{DULCE}\ \underbrace{\text{TIPICO DE}}_{\text{context}}$$

Mikolov et al., 2013



Vector de entrada

(Input vector)

Capa oculta

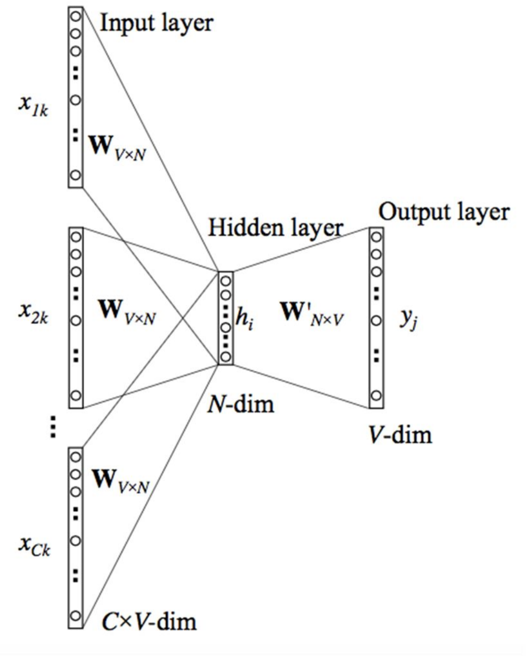
(Hidden layer)

Vector de salida

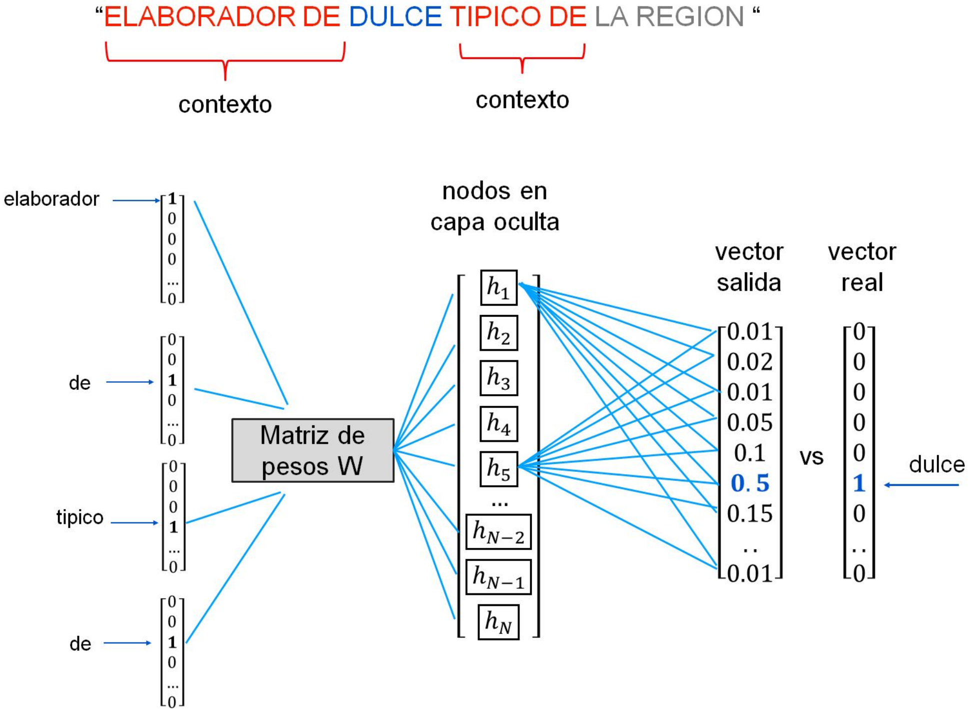
(Output vector)

$$\text{Dulce} \rightarrow \begin{bmatrix}0\\0\\0\\0\\0\\\cdots\\0\\1\\0\\0\\0\end{bmatrix}$$

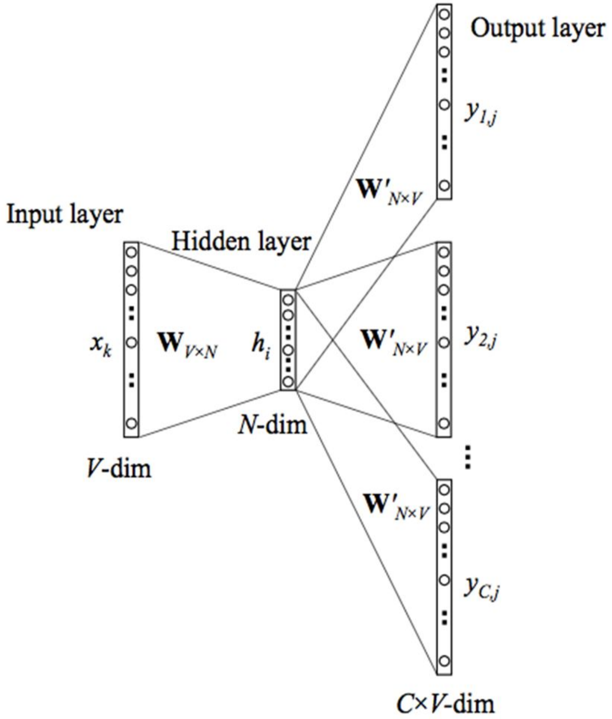
CBOW: Continuous Bag-of-Words

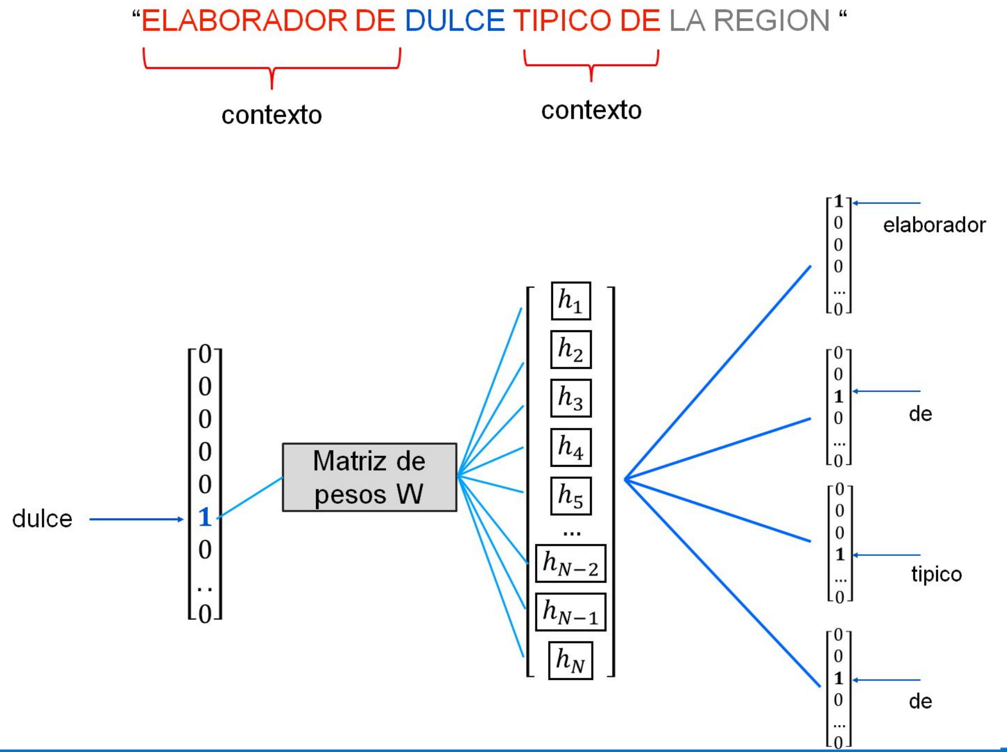


Fuente:http://www.stokastik.in/understanding-word-vectors-and-word2vec/



Skip Gram





“ELABORADOR DE DULCE TIPICO DE LA REGION “

contexto

contexto

Processing with TF iDF

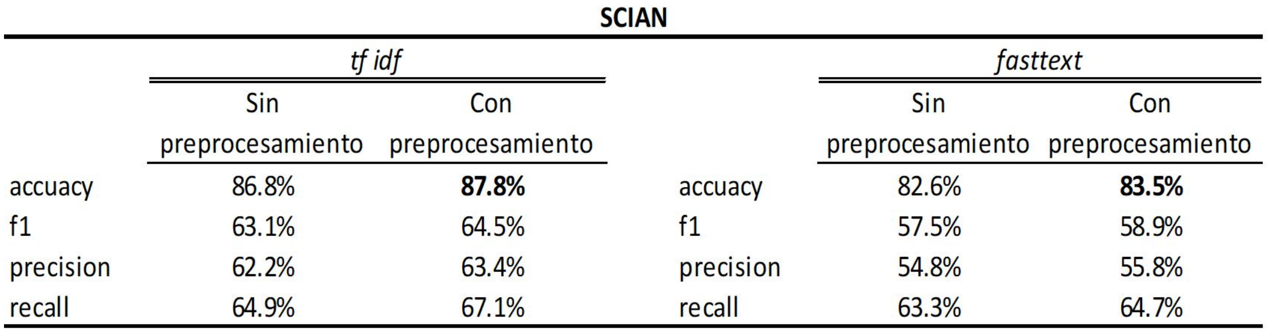
Two TF iDF matrices (each one of 30 000 columns)

- 2-word sequence
- 6-letter sequence

➡

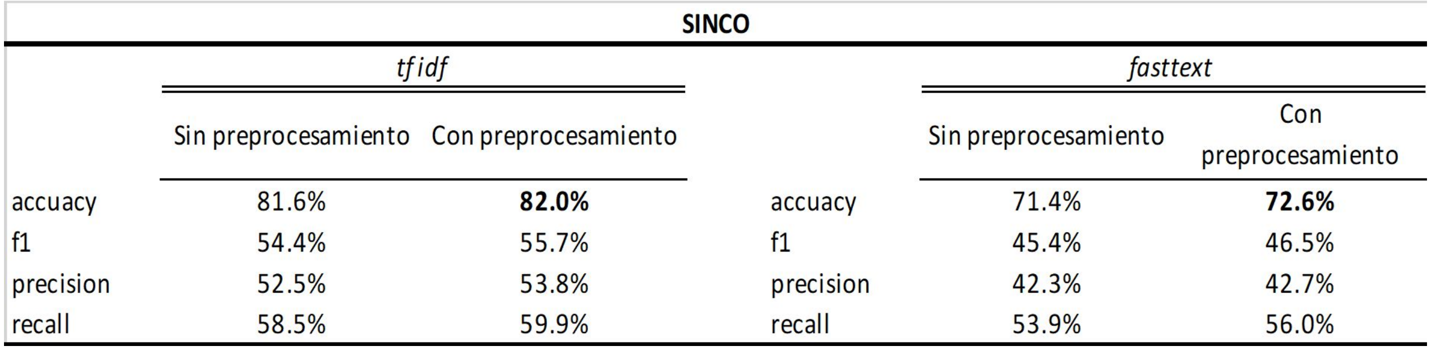
Concatenate the two matrices into one of 60,000 columns + 13 auxiliary variables

Results obtained with TF iDF



SCIAN

	tf idf			*fasttext*	
	Sin preprocesamiento	Con preprocesamiento		Sin preprocesamiento	Con preprocesamiento
accuacy	86.8%	**87.8%**	accuacy	82.6%	**83.5%**
f1	63.1%	64.5%	f1	57.5%	58.9%
precision	62.2%	63.4%	precision	54.8%	55.8%
recall	64.9%	67.1%	recall	63.3%	64.7%



SINCO

	tf idf			*fasttext*	
	Sin preprocesamiento	Con preprocesamiento		Sin preprocesamiento	Con preprocesamiento
accuacy	81.6%	**82.0%**	accuacy	71.4%	**72.6%**
f1	54.4%	55.7%	f1	45.4%	46.5%
precision	52.5%	53.8%	precision	42.3%	42.7%
recall	58.5%	59.9%	recall	53.9%	56.0%

SVM

	Economic Activity	Occupation
6-grams	0.8782	0.8204
6-grams, 10-grams	0.8781	0.8189
6-grams, 2-words	0.8793	0.8188

60 000 - column matrices

Ensamble

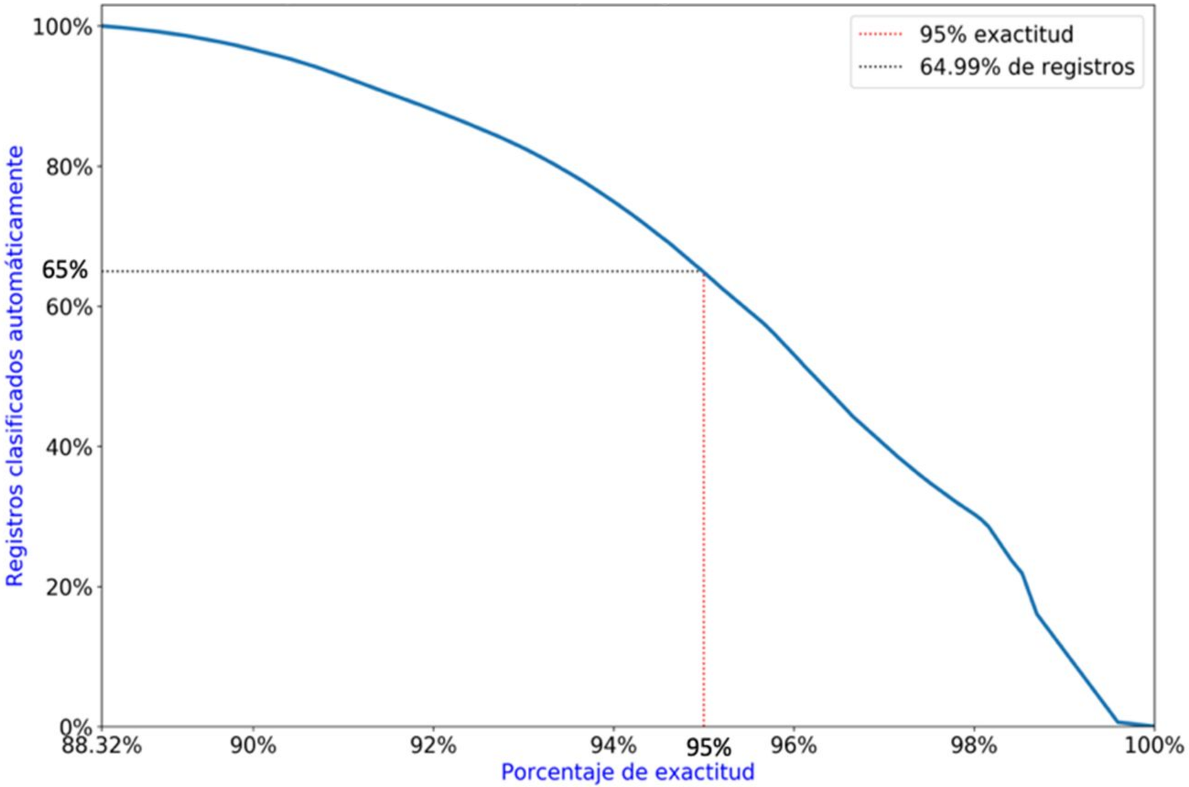
SVM
Logistic regression
Random Forest
Neural Networks
XGBoost
K-NN

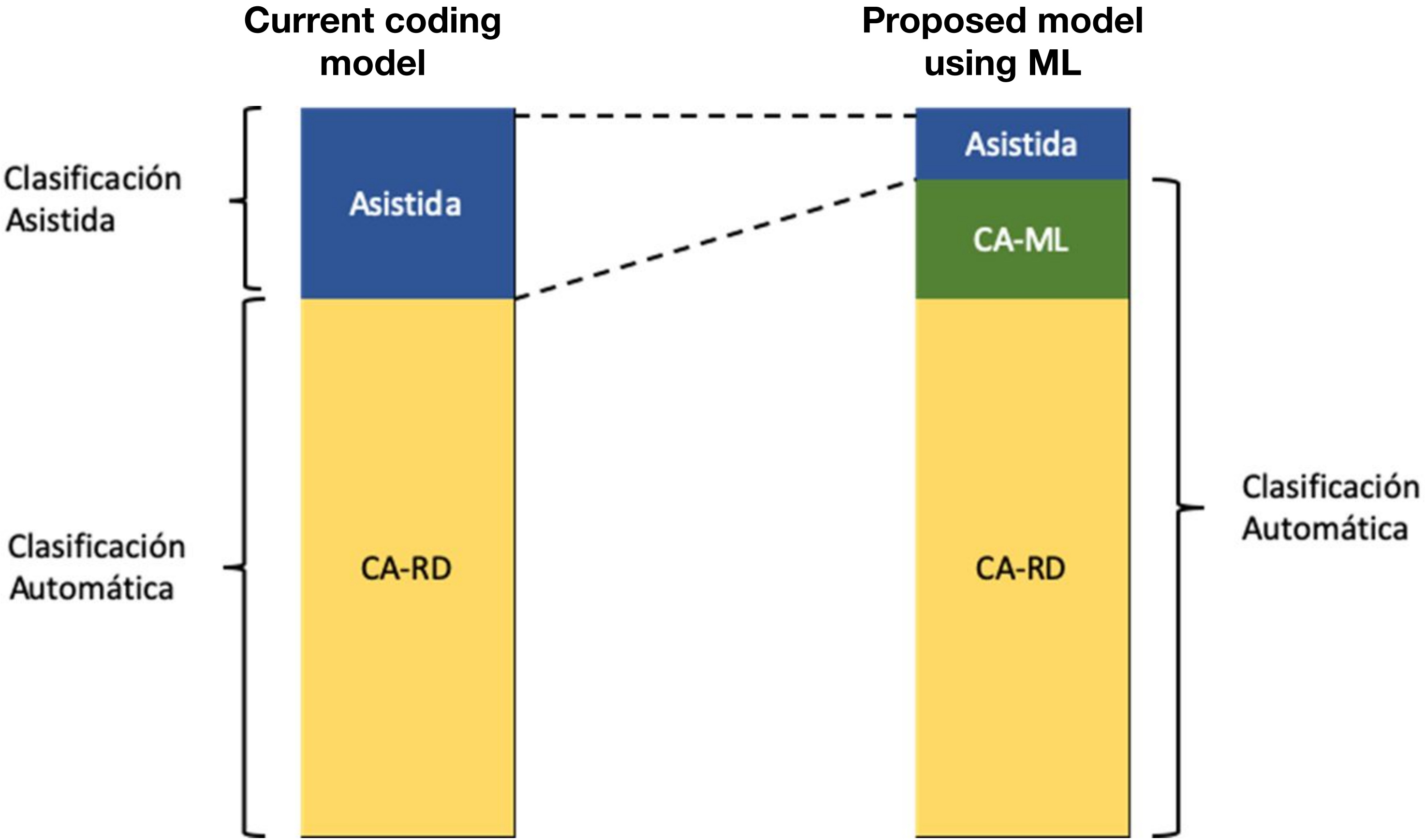
	Economic Activity	Occupation
6-grams	0.8849	0.8474
6-grams, 10-grams	0.8825	0.8647
6-grams, 2-words	0.8905	0.8505

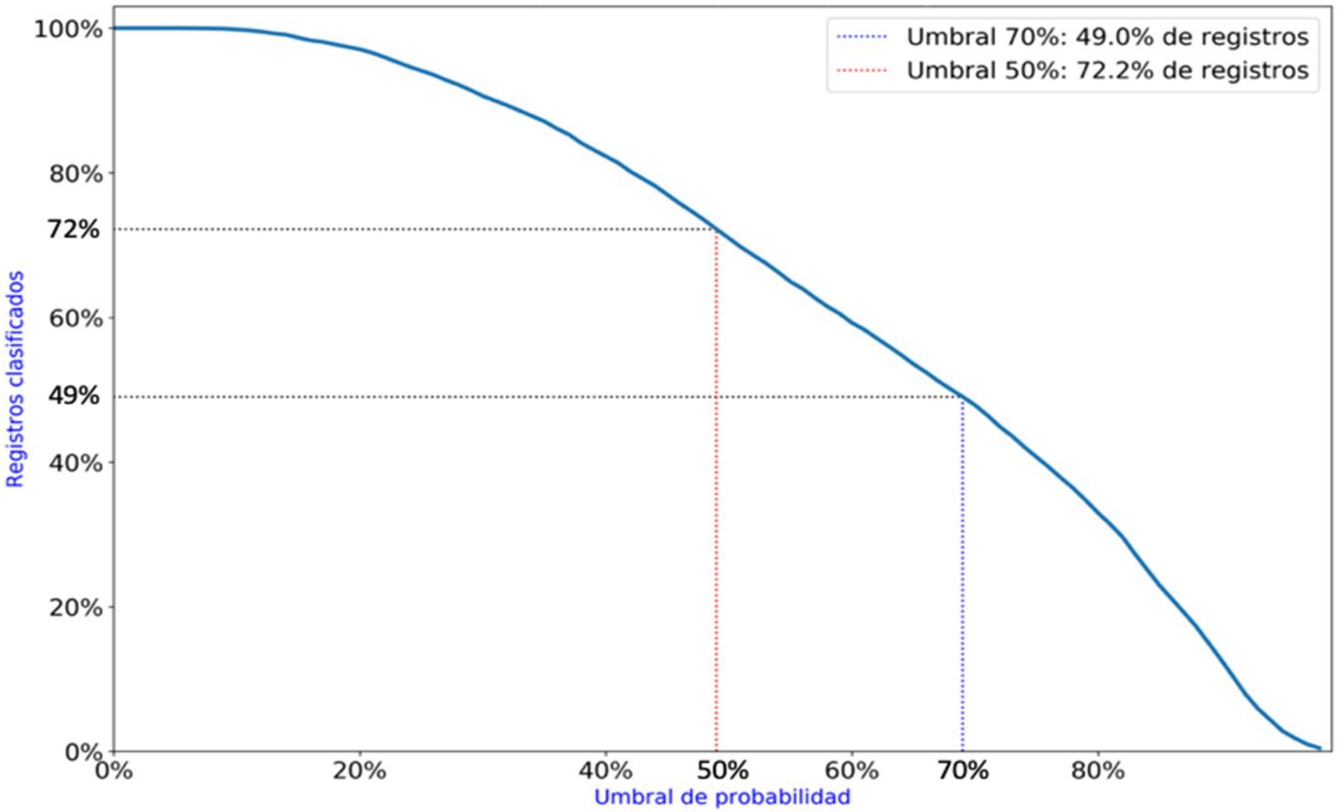
Economic Activity				
	Accuracy	Precision	Recall	F1
Assembly with same weights	0.8905	0.6925	0.6149	0.6365
Assembly with differentiated weights	0.8921	0.6767	0.6420	0.6512

Occupation				
	Accuracy	Precision	Recall	F1
Assembly with same weights	0.8447	0.6441	0.5384	0.5639
Assembly with differentiated weights	0.8505	0.6437	0.5637	0.5831

Trade-off between Percentage of records and Accuracy




Current coding model
Proposed model using ML
Clasificación Asistida
Asistida
Clasificación Automática
CA-RD
Asistida
CA-ML
CA-RD
Clasificación Automática



- Take advantage of the 'probability' metric associated with ML algorithms
- Only those records that pass a probability threshold would be classified with ML
- Determine the threshold

Results of the ML models in Population and Housing Census 2020

How was it done for the productive data of the Census?

- From the already coded Census data, 1 million records were selected through a sampling that was proportional to the size of the main group of each classification and also as a control variable, the coding strategy by which it was coded
- Records that were human-coded were included
- 750 thousand records were taken as a training set and 250 thousand as a test set

OCCUPATION

Where to make the cut-off point?

% OF CODING	ERROR RATE
82.39%	3.87%

→ Certainty ≥ **0.7**

% OF CODING	ERROR RATE
73.44%	2.56%

→ Certainty ≥ **0.8**

% OF CODING	ERROR RATE
55.14%	1.29%

→ Certainty ≥ **0.9**

TRUE AND FALSE WITH CERTAINTY ≥ 0.7						
DIVISION	TRUE	%	FALSE	%	% OF CODING	ERROR RATE
0	851	62.44%	91	6.68%	69.11%	9.66%
1	1,015	41.06%	82	3.32%	44.38%	7.47%
2	24,178	78.21%	676	2.19%	80.39%	2.72%
3	7,728	78.99%	339	3.46%	82.45%	4.20%
4	19,601	73.56%	1,043	3.91%	77.47%	5.05%
5	13,979	72.67%	1,139	5.92%	78.60%	7.53%
6	57,711	87.04%	1,838	2.77%	89.81%	3.09%
7	32,715	88.58%	649	1.76%	90.34%	1.95%
8	10,325	76.00%	441	3.25%	79.25%	4.10%
9	29,971	69.97%	1,666	3.89%	73.85%	5.27%
TOTAL	198,074	79.21%	7,964	3.18%	82.39%	3.87%

INEGI

RESULTS BY DIVISION

ECONOMIC ACTIVITY

Where to make the cut-off point?

% OF CODING	ERROR RATE
86.01%	4%

→ Certainty ≥ **0.7**

% OF CODING	ERROR RATE
78.56%	2.71%

→ Certainty ≥ **0.8**

% OF CODING	ERROR RATE
64.08%	1.55%

→ Certainty ≥ **0.9**

TRUE AND FALSE WITH CERTAINTY ≥ 0.7						
SECTOR	TRUE	%	FALSE	%	% OF CODING	ERROR RATE
10	841	67.17%	58	4.63%	71.81%	6.45%
11	74,475	94.24%	1,401	1.77%	96.01%	1.85%
21	1,107	74.95%	48	3.25%	78.20%	4.16%
22	599	71.39%	39	4.65%	76.04%	6.11%
23	17,326	82.35%	798	3.79%	86.14%	4.40%
31	16,735	84.07%	676	3.40%	87.47%	3.88%
32	4,077	74.42%	187	3.41%	77.84%	4.39%
33	9,059	77.42%	401	3.43%	80.85%	4.24%
43	2,394	56.80%	323	7.66%	64.46%	11.89%
46	28,961	76.57%	1,589	4.20%	80.77%	5.20%
48	7,441	84.73%	172	1.96%	86.69%	2.26%
49	1,181	56.48%	159	7.60%	64.08%	11.87%
51	667	69.48%	37	3.85%	73.33%	5.26%

... Continued

RESULTS BY SECTOR

TRUE AND FALSE WITH CERTAINTY ≥ 0.7

SECTOR	TRUE	%	FALSE	%	% OF CODING	ERROR RATE
52	1,543	82.96%	67	3.60%	86.56%	4.16%
53	639	65.14%	39	3.98%	69.11%	5.75%
54	3,104	76.98%	104	2.58%	79.56%	3.24%
55	0	0.00%	0	0.00%	0.00%	0.00%
56	3,757	69.65%	199	3.69%	73.34%	5.03%
61	10,027	83.06%	359	2.97%	86.03%	3.46%
62	4,824	80.53%	167	2.79%	83.32%	3.35%
71	1,366	76.10%	44	2.45%	78.55%	3.12%
72	11,787	63.91%	1,803	9.78%	73.68%	13.27%
81	14,525	85.60%	347	2.05%	87.65%	2.33%
93	0	0.00%	0	0.00%	0.00%	0.00%
99	0	0.00%	0	0.00%	0.00%	0.00%
TOTAL	216,435	82.57%	9,017	3.44%	82.57%	4.00%

RESULTS BY SECTOR

WITH CERTAINTY ≥ **0.7**

Third coding test - ENIGH

OCCUPATION	
% OF CODING	ERROR RATE
85.72%	3.92%

% ERROR
2.14%

ECONOMIC ACTIVITY	
% OF CODING	ERROR RATE
89.35%	3.55%

% ERROR
1.79%

Population and Housing Census 2020

OCCUPATION	
% OF CODING	ERROR RATE
82.39%	3.87%

ACTIVIDAD ECONÓMICA	
% OF CODING	ERROR RATE
86.01%	4%

Conociendo

México

01 800 111 46 34

www.inegi.org.mx

atencion.usuarios@inegi.org.mx

INEGI Informa