Context-Aware Stereotype Detection:

Conversational Thread Analysis on BERT-based Models

Pol Pastells, Wolfgang S. Schmeisser-Nieto, Simona Frenda, Mariona Taulé

WHY Human annotators need context to label a text as containing stereotypes.

WHAT Do discursive contexts also play a role to interpret a message in NLP?

HOW Fine-tune classification models to detect stereotypes related to immigrants in Spanish. We add different contexts after [SEP] token of the models:

[CLS] + TEXT TO CLASSIFY + [SEP] + CONTEXT

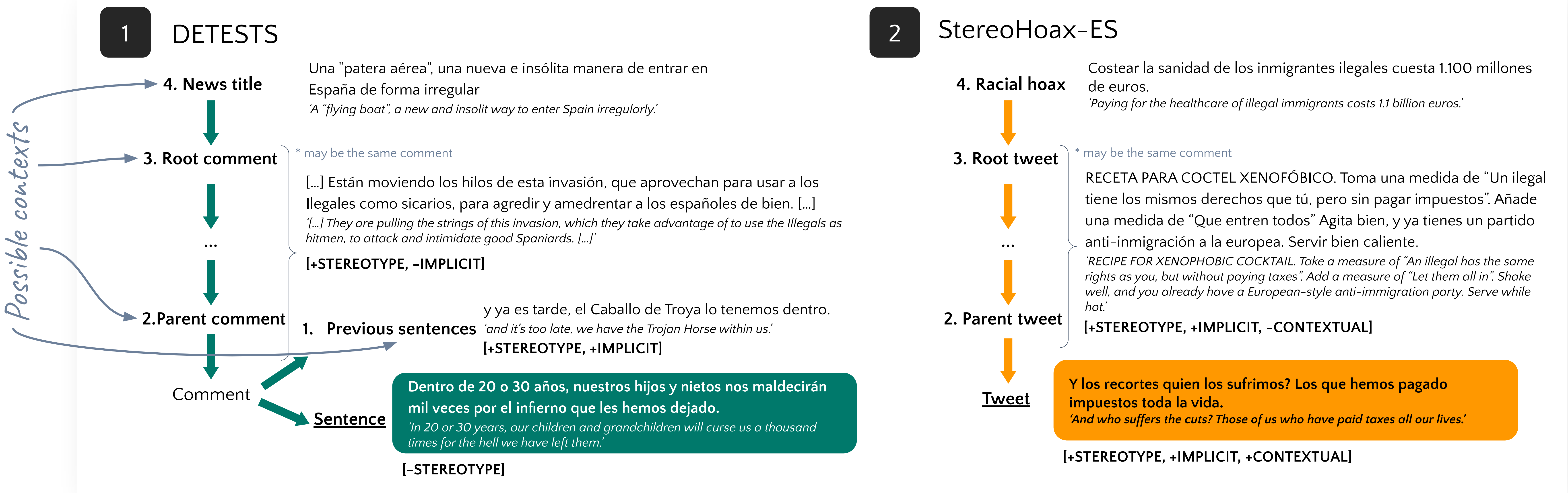
Labels		Nº texts DETESTS	Nº texts StereoHoax-ES
Stereotype	Contextual	-	590
	Explicit	303	1,260
	Implicit	1,056	344
	Total	1,359	1,604
No Stereotype		4,270	3,745
Total		5,629	5,349

MODELS BETO (dccuchile/bert-base-spanish-wwm-cased)

MarIA (PlanTL-GOB-ES/roberta-base-bne)

M-BERT (google-bert/bert-base-multilingual-cased)

Datasets



1 DETESTS

Possible contexts

4. News title

Una "patera aérea", una nueva e insólita manera de entrar en España de forma irregular
'A "flying boat", a new and insolit way to enter Spain irregularly.'

3. Root comment

* may be the same comment

[...] Están moviendo los hilos de esta invasión, que aprovechan para usar a los Ilegales como sicarios, para agredir y amedrentar a los españoles de bien. [...]
'[...] They are pulling the strings of this invasion, which they take advantage of to use the Illegals as hitmen, to attack and intimidate good Spaniards. [...]'

[+STEREOTYPE, –IMPLICIT]

...

2.Parent comment

1. Previous sentences

y ya es tarde, el Caballo de Troya lo tenemos dentro.
'and it's too late, we have the Trojan Horse within us.'

[+STEREOTYPE, +IMPLICIT]

Comment

Sentence

Dentro de 20 o 30 años, nuestros hijos y nietos nos maldecirán mil veces por el infierno que les hemos dejado.
'In 20 or 30 years, our children and grandchildren will curse us a thousand times for the hell we have left them.'

[–STEREOTYPE]

2 StereoHoax-ES

4. Racial hoax

Costear la sanidad de los inmigrantes ilegales cuesta 1.100 millones de euros.
'Paying for the healthcare of illegal immigrants costs 1.1 billion euros.'

3. Root tweet

* may be the same comment

RECETA PARA COCTEL XENOFÓBICO. Toma una medida de "Un ilegal tiene los mismos derechos que tú, pero sin pagar impuestos". Añade una medida de "Que entren todos" Agita bien, y ya tienes un partido anti-inmigración a la europea. Servir bien caliente.
'RECIPE FOR XENOPHOBIC COCKTAIL. Take a measure of "An illegal has the same rights as you, but without paying taxes". Add a measure of "Let them all in". Shake well, and you already have a European-style anti-immigration party. Serve while hot.'

[+STEREOTYPE, +IMPLICIT, –CONTEXTUAL]

...

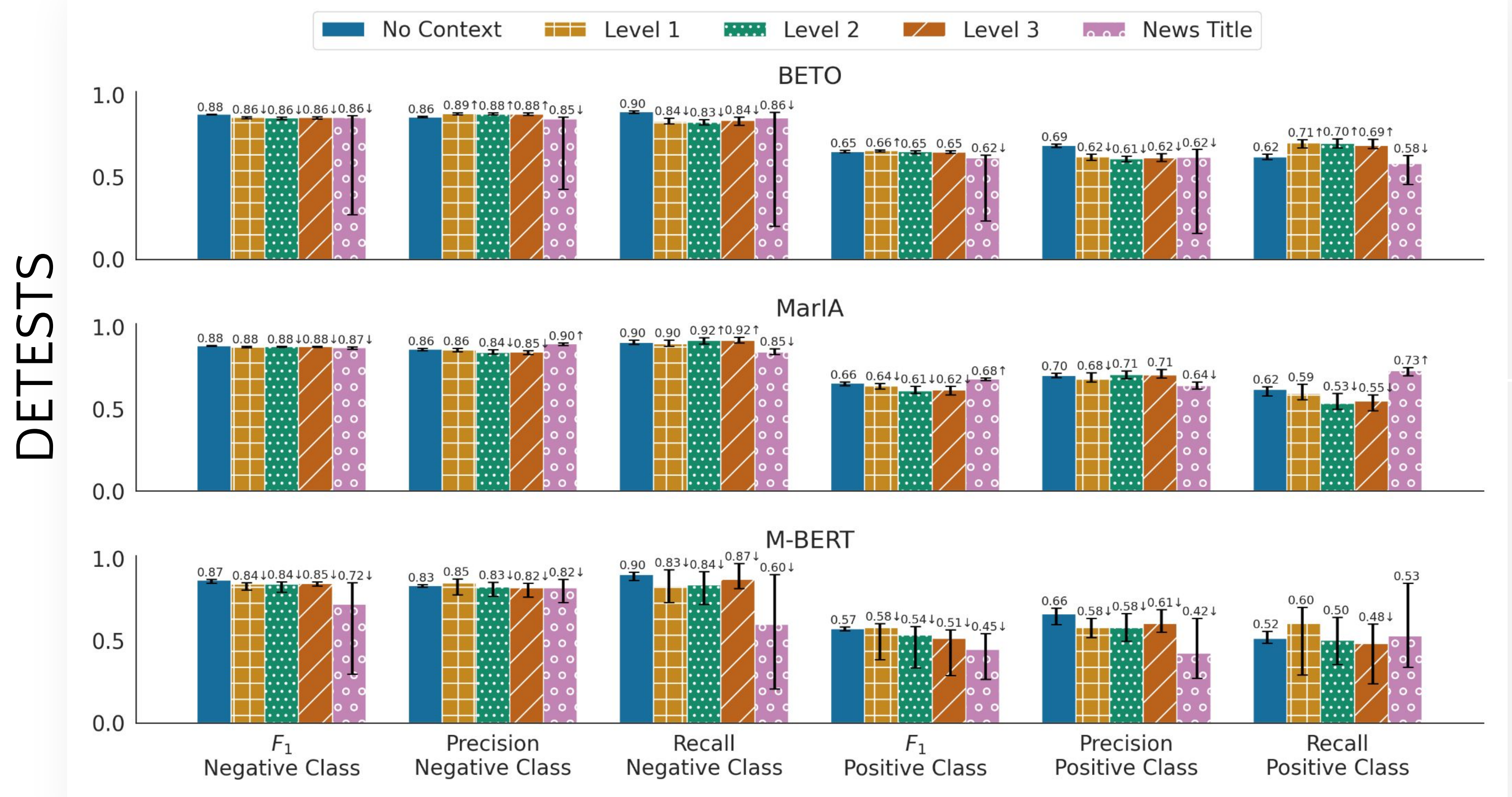
2. Parent tweet

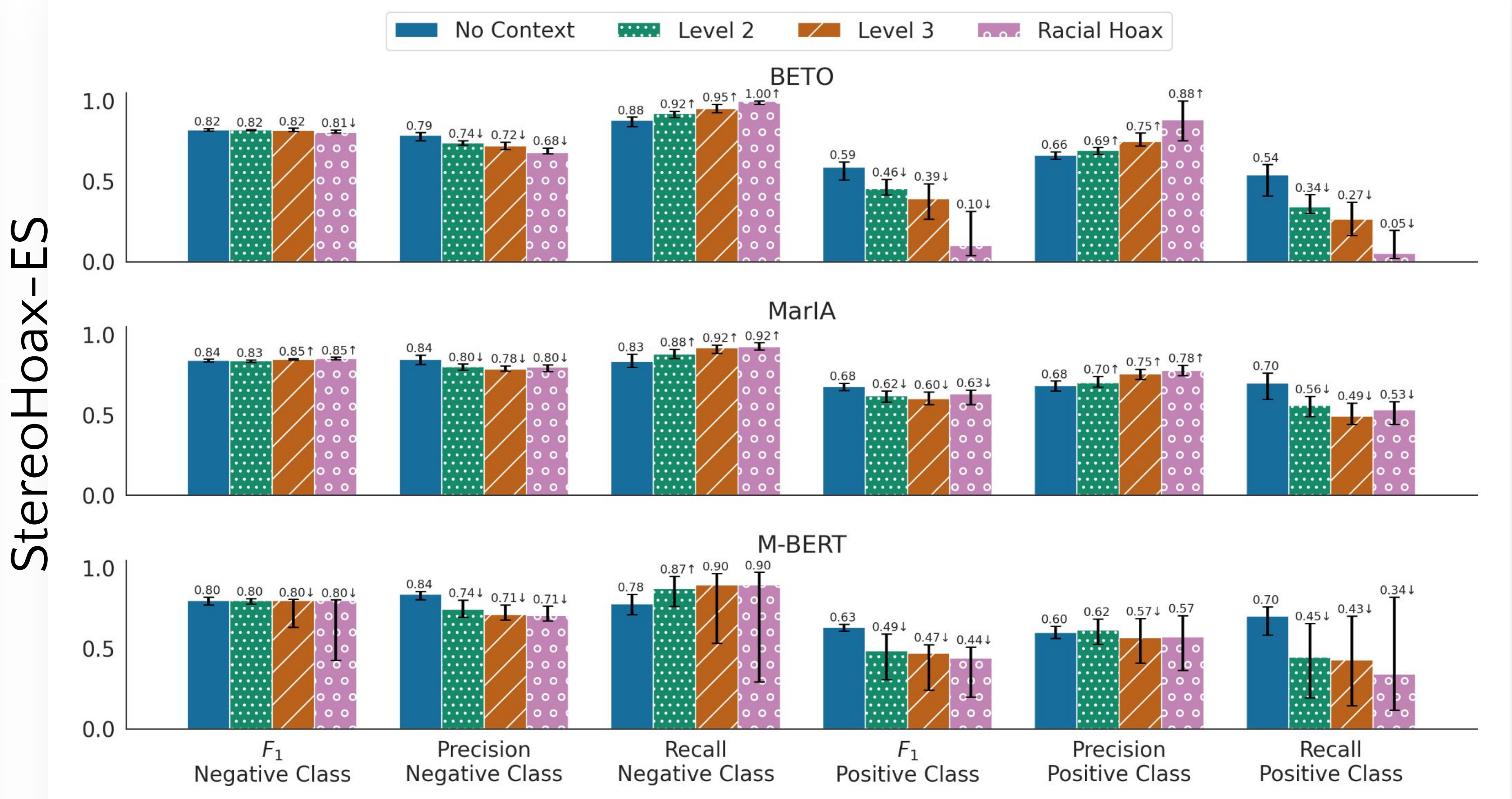
Tweet

Y los recortes quien los sufrimos? Los que hemos pagado impuestos toda la vida.
'And who suffers the cuts? Those of us who have paid taxes all our lives.'

[+STEREOTYPE, +IMPLICIT, +CONTEXTUAL]

Results





Model	Category	No Context > 65% seeds	Level 1 Changes		Level 2 Changes		Level 3 Changes		Level 4 Changes	
BETO	FP	163	**40**	**(25% ↓)**	**50**	**(31% ↓)**	**37**	**(23% ↓)**	4	(2% ↓)
	TN	1173	0		1	(0% ↑)	2	(0% ↑)	3	(0% ↑)
	FN	112	0		1	(1% ↓)	3	(3% ↓)	3	(3% ↓)
	TP	292	16	(5% ↑)	19	(7% ↑)	14	(5% ↑)	2	(1% ↑)
MarIA	FP	169	11	(7% ↓)	7	(4% ↓)	7	(4% ↓)	**41**	**(24% ↓)**
	TN	1187	12	(1% ↑)	14	(1% ↑)	16	(1% ↑)	1	(0% ↑)
	FN	98	8	(8% ↓)	**21**	**(21% ↓)**	**28**	**(29% ↓)**	0	
	TP	277	6	(2% ↑)	3	(1% ↑)	3	(1% ↑)	**35**	**(13% ↑)**
M-BERT	FP	198	6	(3% ↓)	1	(1% ↓)	0		3	(2% ↓)
	TN	1167	0		0		2	(0% ↑)	0	
	FN	86	1	(1% ↓)	1	(1% ↓)	**11**	**(13% ↓)**	0	
	TP	223	3	(1% ↑)	0		0		1	(0% ↑)

- ❖ BETO does **worse** with contexts (level 1, 2 and 3), with more False Positives.
- ❖ MarIA **worsens** with context level 2 and 3 with more False Negatives. Level 4 has a positive bias.

Model	Category	No Context > 65% seeds	Level 2 Changes		Level 3 Changes		Level 4 Changes	
BETO	FP	142	12	(8% ↓)	0		0	
	TN	609	28	(5% ↑)	34	(6% ↑)	55	(9% ↑)
	FN	58	**29**	**(50% ↓)**	**51**	**(88% ↓)**	**113**	**(195% ↓)**
	TP	148	2	(1% ↑)	1	(1% ↑)	0	
MarIA	FP	83	1	(1% ↓)	1	(1% ↓)	0	
	TN	566	11	(2% ↑)	18	(3% ↑)	23	(4% ↑)
	FN	68	**22**	**(32% ↓)**	**22**	**(32% ↓)**	**27**	**(40% ↓)**
	TP	226	3	(1% ↑)	0		0	
M-BERT	FP	80	0		0		0	
	TN	504	24	(5% ↑)	34	(7% ↑)	39	(8% ↑)
	FN	112	**30**	**(27% ↓)**	**22**	**(20% ↓)**	**28**	**(25% ↓)**
	TN	227	0		0		0	

- ❖ All 3 models do **worse** with contexts (level 2, 3 and 4), with more False Negatives.

- ❖ No general improvement using contexts after the [SEP] token on BERT-based models.
- ❖ Results were highly dependent on the dataset used.

pastells/Context-Aware-Stereotype-Detection