Human vs Machine Perceptions on Immigration Stereotypes

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

Universitat de Barcelona · CLiC Centre de Llenguatge i Computació · UBICS · aequa.tech · Università di Torino

WHY Low agreement in the detection of stereotypes due to the high subjectivity of the task, especially in implicit stereotypes.

WHAT Shed light on linguistic distinctions in how humans and various models perceive stereotypes.

Label	Fleiss' kappa
Stereotype	0.75
Contextual	0.48
Implicit	0.15

HOW Classification models to detect stereotypes related to immigrants in a Spanish corpus. Traditional gold standard labels, disaggregated annotation of training data, and instance predictions yielded by GPT-4.

RQ1: Under what conditions do the models exhibit low confidence in their predictions?

RQ2: To what extent do the predictions of the models differ from human annotations? Where do these discrepancies manifest most prominently, and what are the characteristics of these textual instances?

Warning: Offensive content	Stereotype			Contextual			Implicit		
Text	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
And I've been paying social security for more than 38 years. If I knew better, I'd have become an illegal.	1	1	1	0	1	1	1	0	1
The immigrant who comes is not qualified, he has another religion, culture and language but aspires for Spain to support him at the expense of our pensioners. Not all of them are criminals but they are overrepresented in our prisons.	1	1	1	0	0	0	0	0	0
Being an immigrant does not mean being Muslim. I emigrated and I do not believe in any religion. Maybe that's why you don't understand it.	0	0	0	–	–	–	–	–	–

Multilingual Stereotype Corpus (Bourgeade et al., 2023)

- **Source:** Twitter posts
- **Topic:** Immigrants
- **Language:** Spanish
- **Total instances:** 5,349

Labels		N° of instances
Stereotype	Contextual	590
	Explicit	1,260
	Implicit	344
	Total	1,604
No Stereotype	Total	3,745

Methods

Fine-tuning **Zero shot**

1 Soft labels for 3 annotators

0	0	0	**0.05**
0	0	1	**0.27**
0	1	1	**0.73**
1	1	1	**0.95**

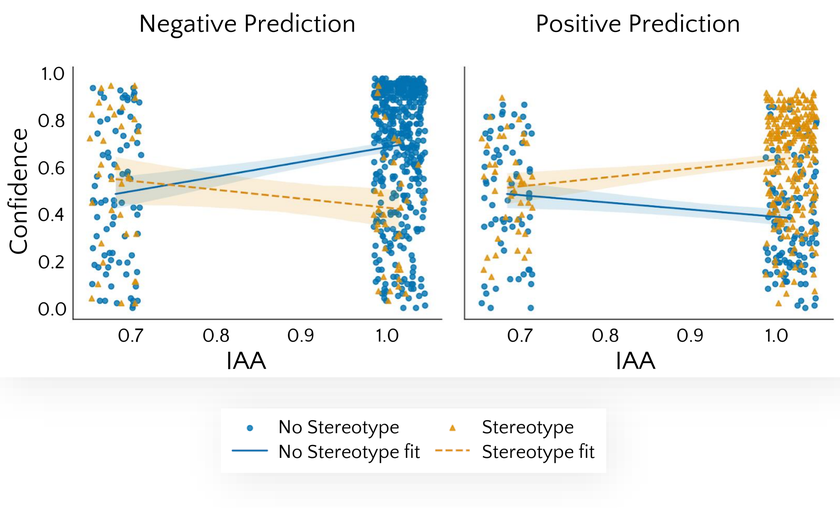
2 Qualitative analysis

How strange 🤦, if they are little angels 🧐

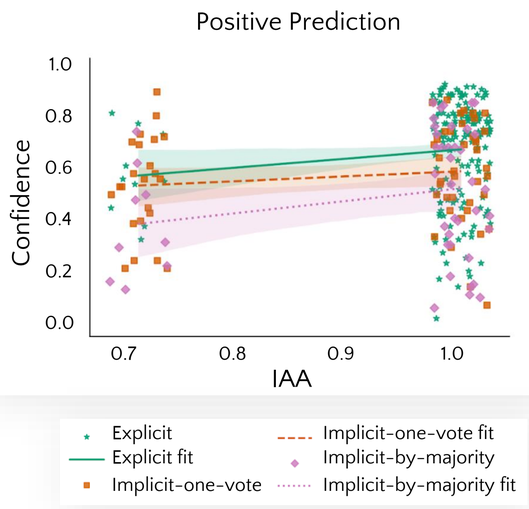
Context needed

Implicit through the use of irony

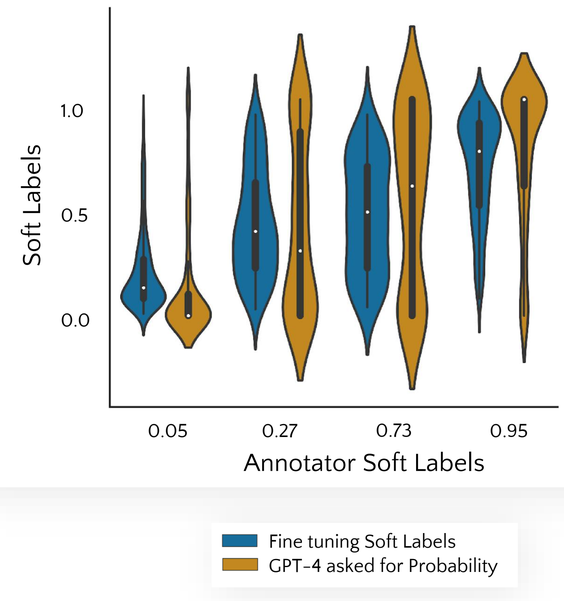
Result 1: Correlation between inter-annotator agreement (IAA) and model confidence.



Result 2: A single *implicit* vote worsens model confidence.



Result 3: Similar results for soft labels



❖**Model confidence** in Spanish stereotype detection aligns with **annotator agreement.**

❖There is **more disagreement** when the texts contain **implicit stereotypes.**

Reference

- Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In Findings of the Association for Computational Linguistics: EACL 2023, pages 686–696, Dubrovnik, Croatia. Association for Computational Linguistics.

Fondazione Compagnia di San Paolo