Human vs. Machine Perceptions on Immigration Stereotypes

Wolfgang S. Schmeisser-Nieto[1,2,*], **Pol Pastells**[1,*], **Simona Frenda**[3,4], **Mariona Taulé**[1,2]
1. Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona, Spain
2. Institute of Complex Systems (UBICS), Universitat de Barcelona, Spain
3. Dipartimento di Informatica, Università degli Studi di Torino, Italy
4. aequa-tech, Turin, Italy
{wolfgang.schmeisser, pol.pastells, matule}@ub.edu
simona.frenda@unito.it

Abstract

Warning: *This paper contains derogatory language that may be offensive to some readers.*
The increasing popularity of natural language processing has led to a race to improve machine learning models that often leaves aside the core study object, the language itself. In this study, we present classification models designed to detect stereotypes related to immigrants, along with both quantitative and qualitative analyses, shedding light on linguistic distinctions in how humans and various models perceive stereotypes. Given the subjective nature of this task, one of the models incorporates the judgments of all annotators by utilizing soft labels. Through a comparative analysis of BERT-based models using both hard and soft labels, along with predictions from GPT-4, we gain a clearer understanding of the linguistic challenges posed by texts containing stereotypes. Our dataset comprises Spanish Twitter posts collected as responses to immigrant-related hoaxes, annotated with binary values indicating the presence of stereotypes, implicitness, and the requirement for conversational context to understand the stereotype. Our findings suggest that both model prediction confidence and inter-annotator agreement are higher for explicit stereotypes, while stereotypes conveyed through irony and other figures of speech prove more challenging to detect than other implicit stereotypes.

Keywords: Stereotype Detection, Immigration, Annotation, Disagreement

1. Introduction

The past decade has seen the rise of social media, and with it, the propagation of misleading information aimed at stigmatizing vulnerable social groups such as women, immigrants and the LGBTQIA+ community, and the increasing spread and reinforcement of stereotypes about these groups. However, this trend has also been accompanied by research in the computational linguistics (CL) community aiming to tackle this phenomenon. Those efforts have focused on creating datasets annotated with stereotypes (Ariza-Casabona et al., 2022; Javier Sánchez-Junquera and Ponzetto, 2021) and other related categories that are helpful for stereotype detection tasks, such as stereotype taxonomies or implicitness, as well as on improving detection and classification techniques (Fokkens et al., 2018; Sap et al., 2020).

A recent paradigm in the CL community is the creation of disaggregated datasets to model subjective phenomena in a *perspectivist* manner, training the models on specific perspectives of a segment of the population or an individual (strong perspectivism) (Cabitza et al., 2023), or training the models considering the disagreement among annotators (soft perspectivism) (Uma et al., 2020). In traditional approaches, annotations are represented as a single value known as the gold standard (aggregated annotation) or hard label, determined through methods such as majority voting. However, opinion tasks, such as the identification of hate speech and stereotypes (Chulvi et al., 2023), tend to generate disagreement among annotators due to the different perceptions of those topics driven by demographic characteristics. In particular, in our work, we employed the computation of soft labels, calculated using softmax normalization as suggested by Uma et al. (2020). Therefore, the predictions of the model will vary in accordance with the inputs it receives, whether they are hard or soft labels.

In this work, we explore how Fine Tuning with Soft Labels (FT-SL) and with Hard Labels (FT-HL) modifies the performance of fine-tuned state-of-the-art large language models, and the predictions obtained with zero-shot learning, namely exploiting GPT-4 (OpenAI, 2023), in the detection of stereotypes in Spanish. In particular, we use a Spanish dataset (Bourgeade et al., 2023), containing reactions from Twitter to false news involving migrants published mainly from 2019 to 2021. This dataset was annotated with the presence and absence of stereotypes regarding immigrants, and whether they were expressed explicitly or implicitly.

Exploiting the disaggregated version of this dataset, we want to investigate how annotators and models recognize stereotypes by looking in particular at their form of expression: explicit and implicit.

*These authors contributed equally to this work

LREC-COLING 2024, pages 8453–8463

To do that, we compare the human annotation and the performance of the before-mentioned models, looking at the models' confidence and agreement among the annotators, as well as the linguistic characteristics of the texts. We hypothesize that implicit stereotypes are more difficult to be recognized by models, and can create more disagreement among humans, because they imply the need for an inference to a secondary or additional meaning. Therefore, we aim to answer the following questions:

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

RQ2 To what extent do the predictions of FT-HL, FT-SL, and GPT-4 differ from human annotations? Where do these discrepancies manifest most prominently, and what are the characteristics of these textual instances?

To answer our RQs, we fine-tuned RoBERTa-BNE (Fandiño et al., 2022) using the traditional method (hard labels), as well as soft labels. We also gave GPT-4 the task of predicting each instance of the test set with the presence or absence of stereotypes. Finally, we performed quantitative and qualitative analyses comparing the predictions of the models and the decisions of the annotators, observing the linguistic characteristics of the texts where they differ also in terms of confidence and agreement.

The contributions of this paper are: (i) to use a dataset in Spanish with non-aggregated annotations containing stereotypes about immigrants to train a model with soft labels; (ii) to present an evaluation of the predictions of the RoBERTa-BNE models and GPT-4, and compare those models with the human annotations, since each model has been pre-trained on different data, and each annotator has different background; (iii) to present a correlation between the confidence of the models in their predictions in contrast to the agreements of the annotators; and (iv) to present a qualitative analysis from a linguistic perspective of the texts where the confidence of the models and the agreement among annotators are very different.

This paper is divided into five sections. Section 2 presents state-of-the-art research on stereotype detection. Section 3 describes the dataset and Section 4 the experimental settings used to test our hypotheses, as well as the results. Section 5 presents an analysis of both our quantitative and qualitative results. Finally, Section 6 discusses conclusions and future work.

2. Related Work

A stereotype has been defined by social psychologists as a cognitive phenomenon consisting of an exaggerated set of beliefs about a social group (Hamilton, 1981). Therefore, individuals perceived as members of a specific social group undergo a process of categorization in which the features associated with that social group are attributed to all of its members (Allport et al., 1954). From a psycholinguistic perspective, stereotypes can be manifested through language, and they can be expressed explicitly and implicitly (Greenwald and Banaji, 1995). Explicit stereotypes convey a direct message, openly displaying the associated attributes, for instance, with the use of pejorative adjectives (Collins and Clément, 2012; D'Errico and Paciello, 2018). In contrast, implicit stereotypes are subtle, indirect and require a process of inference to be interpreted by the reader. There are certain linguistic forms in which implicit stereotypes are conveyed within a text, such as through metaphor and irony (Collins and Clément, 2012), negation (Beukeboom et al., 2010) or entailments (Pettigrew and Meertens, 1995). Recently, there have been attempts to formalize the different strategies for expressing implicitness in stereotypes to formulate clear standardized criteria for annotators (Schmeisser-Nieto et al., 2022).

From a computational perspective, there have been increasing efforts to improve the automatic detection of stereotypes in texts. Sap et al. (2020) proposes a formalism to capture the pragmatic implications of stereotypes that semantic approaches fail to represent. Following the idea that stereotypes are present in narratives, Fokkens et al. (2018) introduces *microportraits*, a collection of descriptions of a target group containing stereotypes referring to Muslims. Card et al. (2016) collects similarities in descriptions of people with to create "latent personas" using unsupervised learning.

Other studies that work with narratives for analyzing stereotypes in conversational contexts include DETOXIS (Taulé et al., 2021), DETESTS (Ariza-Casabona et al., 2022) and NewsCom-TOX (Taulé et al., 2024), that comprise comments from online news; StereoImmigrants (Sánchez-Junquera et al., 2021), derived from speeches on immigration at the Spanish parliament; and a multilingual dataset (Bourgeade et al., 2023) of Twitter posts (tweets) in response to racial hoaxes, which are communicative acts featuring distorted and misleading information in the form of a threat to individuals' or societies' health and safety, in which the protagonist is a person, or a group of people described in terms of their ethnicity, nationality, or religion (Cerase and Santoro, 2018). All the aforementioned datasets provide annotated data for detecting stereotypes regarding immigration in Spanish, which constitute a valuable resource due to the scarcity of such datasets.

Considering the subjectivity of this phenomenon,

we leverage the disagreement among the annotators about the presence of stereotypes in Spanish tweets. We follow the new theoretical framework of soft perspectivism using the learning with disagreement approach (Uma et al., 2020). Other studies have used disagreements differently, with a focus on improving their models. A perspectivist approach has also been used to exclude data with low agreement (Beigman Klebanov and Beigman, 2009), to provide a supplement to the gold standard (Plank et al., 2014; Romberg, 2022), or to completely omit the gold standard to train models with the disagreements without aggregation (Rodrigues and Pereira, 2018; Uma et al., 2020). In our work, we computed the soft labels using the softmax function (Uma et al., 2020). Moreover, we embraced perspectivism not only to investigate how the detection of stereotypes could be improved but also to obtain valuable information on the linguistic characteristics that are more difficult to detect according to the aggregation method. We do so through the analysis of the different predictions given by the models compared to the human annotation.

3. Dataset

The dataset[1] is based on a multilingual corpus (Bourgeade et al., 2023) compiled from tweets. These tweets were part of conversation threads in response to verified hoaxes, collected in French, Italian, and Spanish. For this work, we used the Spanish subset, which comprises 4,751 tweets. 598 additional tweets were selected and subjected to annotation following the same strategies and guidelines. The process involved two main steps: initially, search strategies were used to find tweets related to the hoaxes collected manually from debunking sites; and subsequently, the tweets were filtered using keywords associated with hoaxes referring to immigrants.

The dataset was annotated with the following hierarchical binary labels by three annotators (two linguistics students trained for this task and a researcher): *stereotype* indicates the presence of stereotypes regarding immigrants; only if a stereotype is present, *contextual* refers to the need for conversational context, i.e., previous messages, to interpret it; and *implicit* indicates whether the stereotype (if present) is implicit, requiring the reader to infer and interpret it. The following tweet is an example of an implicit stereotype, wherein the first two annotators marked the presence of the stereotype, while the third annotator tagged its absence:

1. *And I've been paying social security for more than 38 years. If I knew better, I'd have become an illegal.*[2]
[Stereotype:Yes] [Contextual:No] [Implicit:Yes]

Table 1 shows the inter-annotator agreements for the three hierarchical binary labels.

Label	Av. pairwise % Agreement	Fleiss' Kappa
Stereotype	89.34%	0.75
Contextual	89.00%	0.48
Implicit	85.61%	0.15

Table 1: Inter-annotator agreement test.

Out of the 1,604 (30.0%) tweets in the dataset containing stereotypes, 590 (36.8%) require context from previous messages to be interpreted, and 344 (21.4%) are expressed implicitly, according to the majority vote labels.

The dataset was divided, with 80% designated for the training set, further split into training and validation subsets, and 20% allocated for the test set. The subsets were stratified to maintain the same distribution of implicit stereotypes and racial hoax topics in each split. To avoid data leakage, we also separated tweets extracted from different hoaxes into different sets, the training set has 15 hoaxes and the test set has 13. Table 2 shows a summary of the distribution of the labels.

4. Models' Predictions

To establish a baseline, we initiated our analysis using the unigram term frequency-inverse document frequency (TFIDF) representation in conjunction with a Support Vector classifier (SVC). This initial approach served as a reference point for evaluating the effectiveness of more advanced techniques.

We investigated the efficacy of different transformer-based models from the BERT family for stereotype detection in Spanish with both hard and soft labels. The selected models, obtained from the *Huggingface* transformers library (`https://huggingface.co/`), were: (i) *dccuchile/bert-base-spanish-wwm-cased*. BETO (Cañete et al., 2020), which has a similar size to BERT-Base and was trained with the Whole Word Masking technique. (ii) *PlanTL-GOB-ES/roberta-base-bne*. RoBERTa-BNE (Fandiño et al., 2022), which is based on the RoBERTa base model. It was pre-trained using the largest Spanish corpus known to date. (iii) *pysentimiento/robertuito-base-uncased*. RoBERTuito (Pérez et al., 2022), which was trained following RoBERTa guidelines on 500 million tweets.

[1] The dataset is available upon request.

[2] Examples have been translated into English to guarantee anonymity.

Subsets	With Stereotypes	Without Stereotypes	Context Needed	Implicit	Explicit	Total
Training set	1,246	3,031	465	270	976	4,277
Test set	358	714	125	74	284	1,072

Table 2: Label distribution among the training and test sets.

RoBERTa-BNE proved to be the best model (see Figures 1 and 2), and is the one we used for the analyses. Building on recent advancements in NLP, we drew inspiration from the work of (Nityasya et al., 2023) and incorporated a few-shot baseline with the same conditions as our main model. Specifically, we trained a RoBERTa-BNE model with a single epoch, employing a minibatch strategy comprising 10 positive and 10 negative cases.

In addition to these models, we also explored a zero-shot approach to stereotype detection using GPT-4. Considering the primary objective of this work, i.e., evaluation of the performance of the models, with a particular focus on the probability and confidence of their predictions, Table 3 shows their results in terms of the F_1 score for the positive and the negative classes, along with accuracy, and report the confidence of models trained on hard labels as median and standard deviation of the distribution of predictions obtained in the test set, while the probability of the predictions of the other models is reported in Table 4.

Model	F_1 pos	F_1 neg	Acc	Confidence
TFIDF+SVC	0.48	0.83	0.75	
FT-HL Few-shot	0.22	0.73	0.60	0.01±0.01
FT-HL	0.69	0.80	0.76	0.64±0.27
FT-SL	0.71	**0.86**	0.81	
GPT-4	**0.72**	**0.86**	**0.81**	
GPT-4P	0.72	0.84	0.80	

Table 3: F_1 score for the positive class (containing stereotypes) and negative classes, and overall accuracy score across all models. Confidence is shown for the fine-tuned models with hard labels. FT-HL and FT-SL stand for the RoBERTA-BNE model fine-tuned with hard and soft labels, respectively; GPT-4P is the GPT-4 model asked for the probability of a stereotype.

Fine-tuning with hard labels For fine-tuning the models with hard labels, we used a learning rate of 10^{-5} with a cosine scheduler set with 20 epochs as the total number of training steps. We trained in batches of 32 tweets, checking the results every 10 steps, with early stopping. Figure 1 shows the training and validation losses. We named the best resulting model, RoBERTa-BNE fine-tuned with hard labels, FT-HL. Table 3 shows that this model achieves good performance compared to the baselines.

We also quantified the confidence of the prediction, by re-scaling the softmax probabilities, using the following formula citepcasola2023confidence:

$$\text{conf}(x) = \frac{\text{Max1}(x) - \text{Max2}(x)}{|\text{Max1}(x) + \text{Max2}(x)|}, \quad (1)$$

where $\text{Max1}(x)$ and $\text{Max2}(x)$ represent the largest and second-largest probability values, respectively, obtained from the predicted logits softmax.

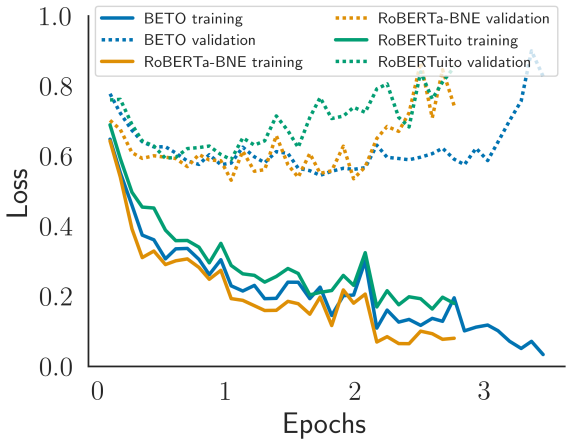

Figure 1: Training and validation loss for the BETO, RoBERTa-BNE and RoBERTuito models using hard labels.

Fine-tuning with Soft Labels Given some data $\{x_i, y_i\}$ with unaggregated annotations, we estimate the human label distribution $p_{hum}(y \mid x)$, i.e., the soft labels, according to a softmax normalization (Uma et al., 2020):

$$p_{hum}(y_i = j \mid x_i) = \frac{exp(d_i^j)}{\sum_a exp(d_i^a)}, \quad (2)$$

where d_i^j is the number of times the annotators chose the j-th class for the i-th data point.

Therefore, the annotator-assigned soft labels encompass four distinct values: 0.05 when none of the annotators label a tweet as containing stereotypes, 0.27 when only one annotator identifies stereotypes, 0.73 when two annotators concur on the presence of stereotypes, and 0.95 when all three annotators unanimously agree.

Model	Cross Entropy	Correlation	Std pos	Std neg
FT-SL	**0.27**	0.62	0.32	0.26
GPT-4P	1.44	**0.64**	0.38	0.21

Table 4: Cross-entropy score, Pearson correlation and standard deviation for positive and negative predictions across models with soft labels.

To predict the soft labels with the different BERT models, we fine-tuned them for a regression task, instead of a classification task for hard labels. For the soft-label models, we set the learning rate at 2×10^{-5}, but otherwise used the same setup as with hard labels. Figure 2 shows the training and validation losses. We named the best resulting model, RoBERTa-BNE fine-tuned with soft labels, FT-SL.

The use of soft labels improves the hard metrics (see Table 3) while maintaining the different views of all the annotators. Table 4 shows the cross-entropy and Pearson correlation coefficient for the models with soft labels. Note that the minimum possible value for the cross-entropy is 0.15, which is equal to the annotator soft-label entropy. We see that FT-SL achieves the best cross-entropy score.

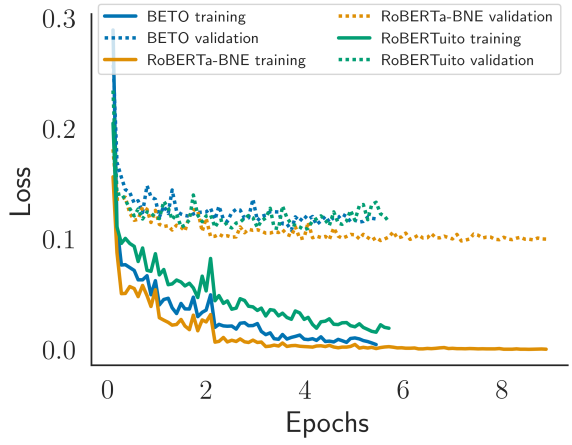

Figure 2: Training and validation loss for the BETO, RoBERTa-BNE and RoBERTuito models using soft labels.

Prediction of Stereotypes with GPT-4 We took a zero-shot approach with GPT-4. Using the OpenAI API, we first gave the following system message[3] to the model: "You are a linguist with expertise in annotating sentences with stereotypes."

We then added the following content message: "You must classify the sentence in double parentheses according to whether it contains any racist stereotypes or not. Return a single integer without comments: '1' if positive, '0' if negative. ((<sentence>))"

To compare with the soft labels, we asked instead for the probability of stereotypes (GPT-4P): "You must give the probability that the sentence in double parentheses contains any racist stereotypes. Return a single real number between 0 and 1 without comments. ((<sentence>))"

Tables 3 and 4 show the results of these prompts. GPT-4 achieves the best overall scores with hard labels, tying with FT-SL on the F_1 negative class score. When asked for probabilities (GPT-4P), it achieves the highest correlation, although with a poor cross-entropy score.

5. Human vs. Machine Analysis

In this section, we compare the confidence and performance of the models FT-HL, FT-SL and GPT-4P, with the gold standard and soft labels from the human annotators.

5.1. Hard Labels

To evaluate the model with hard labels, we used an IAA test, the instance agreement percentage. In our scenario involving three annotators, the IAA yields one of two values: either total agreement among all annotators ($IAA = 1$) or one annotator disagrees ($IAA = 0.67$).

Figure 3 (top) represents the relationship between the confidence scores of FT-HL and the IAA for both negative and positive predictions. To ensure clarity in the visualization and prevent overlapping data points, we introduced jitter in the x-direction.

False positives are represented as blue squares on the right-hand of the figure, while false negatives are denoted by orange triangles on the left-hand. An examination of these plots revealed several noteworthy observations. We found that for the true positives and true negatives, the model tends to exhibit higher confidence levels when all three annotators agree on the presence of stereotypes. This result reinforces the reliability of the model in cases where there is a clear consensus among human annotators. Conversely, when the model predictions are incorrect, its confidence tends to deviate further from the IAA. This deviation indicates a decrease in the performance of the model and highlights the difficulty it faces when handling instances where human annotators do not concur.

Figure 3 (bottom) represents an analysis of tweets containing stereotypes. Using majority voting, stereotypes can be categorized as either explicit or implicit. We took into account the disaggregated annotations of the *implicit* label. This approach allowed us to categorize tweets as explicit,

[3]Prompts have been translated from Spanish.

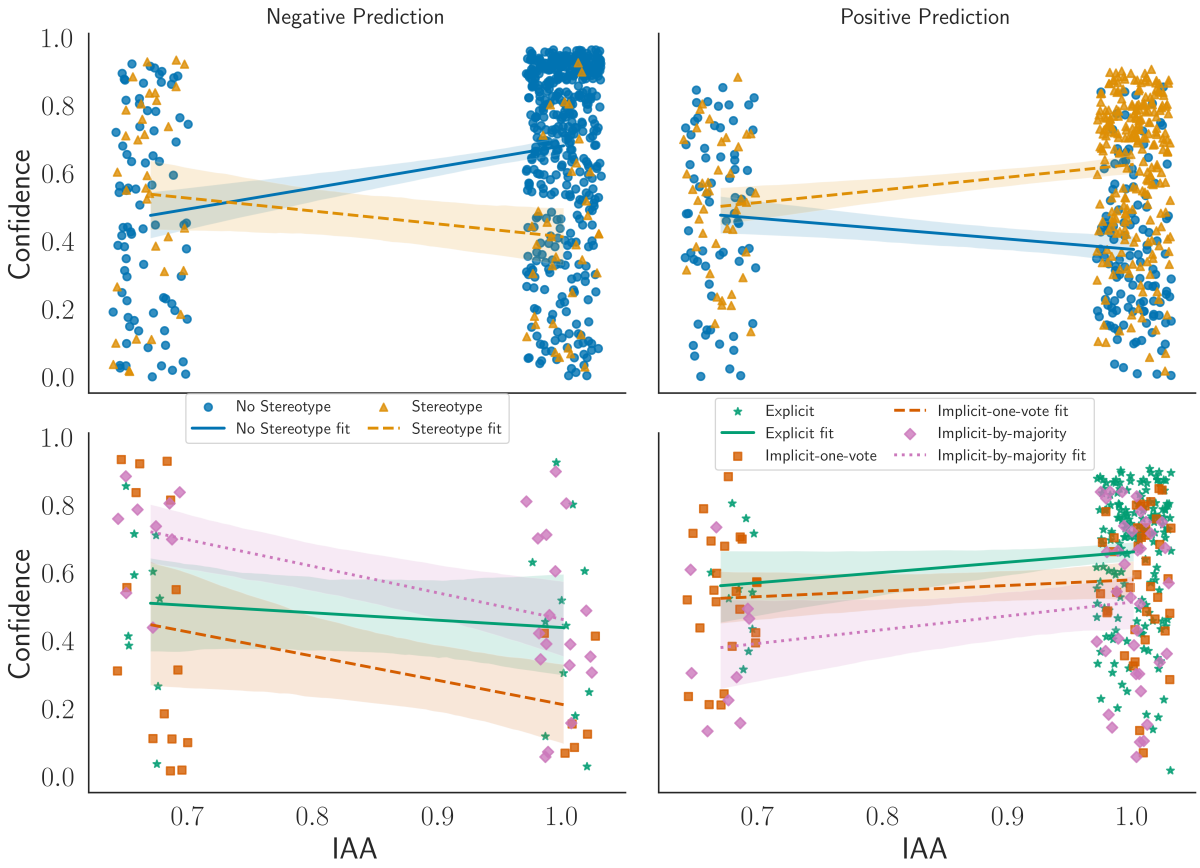

Figure 3: Comparison of FT-HL confidence scores versus IAA, computed as the instance percentage agreement, with negative predictions on the left and positive predictions on the right. (Top) Gold labels for instances without stereotypes are represented as blue circles, and gold labels for instances with stereotypes are represented as orange triangles. (Bottom) Tweets containing stereotypes. Explicit annotations are represented as green stars, implicit-one-vote annotations as red squares, and implicit-by-majority annotations as purple rhomboids. Jitter is added to the x-axis. The straight lines represent a linear regression fit for the data, and the shaded area corresponds to a 95% confidence interval computed with the bootstrapping method.

implicit-one-vote (only one annotator considered it implicit) or implicit-by-majority (two or more annotators considered it implicit).

Our findings revealed patterns in the confidence of the model when handling instances with explicit or implicit stereotypes. When the predictions of FT-HL align with the gold labels, it tends to display lower confidence when dealing with implicit stereotypes than when dealing with explicit ones. This suggests that the model may be less certain when identifying subtle or context-dependent stereotypes.

Conversely, when the model predictions are incorrect (false negatives), the model exhibits more confidence in its predictions involving explicit stereotypes. These findings remained consistent even when accounting for the implicit-one-vote stereotypes, which proved to be more challenging to classify than the explicit stereotypes, but easier than the implicit-by-majority stereotypes.

5.2. Soft Labels and Probability

For the model trained with soft labels and GPT-4P, our analysis involves a direct comparison between the predictions of the models and the soft labels derived from the human annotations.

Figure 4 employs a violin plot to explore the distribution of the predicted soft labels and probability across the four annotator options. Violin plots are useful for depicting variability within different categories. The width of the violin along the y-axis represents the density of tweets falling into each category. It is worth noting that the violin plot displays small values outside the range of 0 to 1. These values are just artifacts of the visualization technique.

Our visual analysis reveals some compelling insights. Notably, the predictions of the model align closely with the original soft labels when all annotators unanimously agree on the presence or absence of stereotypes. This alignment is evident

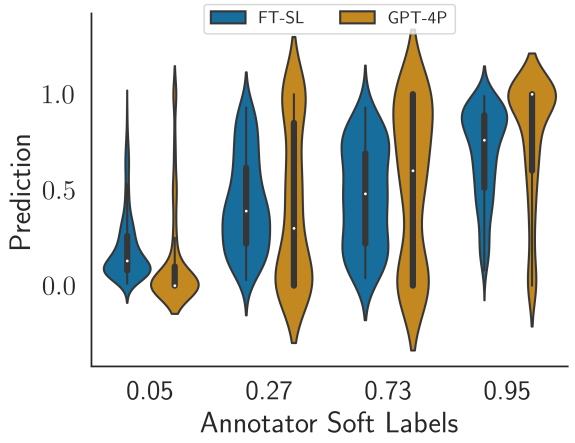

Figure 4: Prediction by FT-SL (soft labels) and GPT-4P (probabilities) versus annotator soft labels.

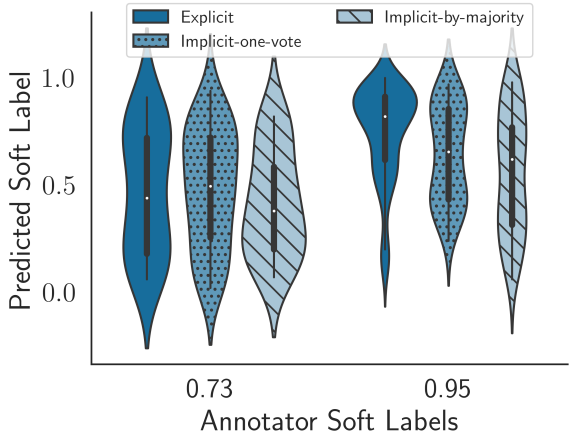

Figure 5: Predicted soft labels by FT-SL versus annotator soft labels for the instances annotated as explicit, implicit-on-vote or implicit-by-majority.

through the more concentrated distributions near the coordinates (0,0) and (1,1) in Figure 4. In other words, both models predictions strongly correlate with the unanimous decisions of the annotators.

Focusing our analysis on stereotypes, we also explore equivalent violin plots in Figure 5 for FT-SL and in Figure 6 for GPT-4P, differentiating between instances with implicit and explicit stereotypes. We see that the explicit stereotypes predicted soft labels are more closely correlated with the annotator soft labels than the implicit stereotypes. Furthermore, as with hard labels, even when a single annotator labeled a tweet as *implicit*, the model performs more poorly than with explicit stereotypes.

5.3. Qualitative Analysis

We based our qualitative analyses of the predictions on the presence of stereotypes made by RoBERTa-BNE models trained with hard (FT-HL)

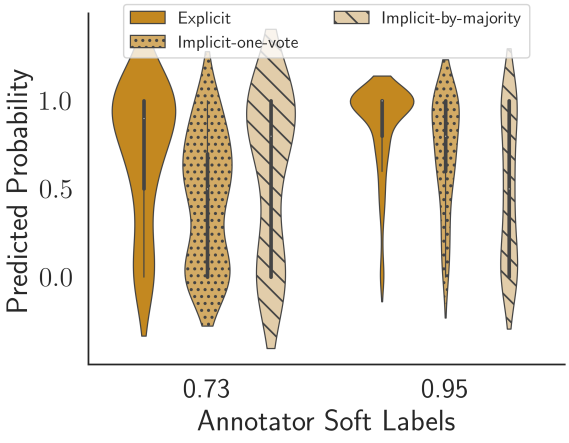

Figure 6: Predicted probabilities by GPT-4P versus annotator soft labels for the instances annotated as explicit, implicit-on-vote or implicit-by-majority.

and soft labels (FT-HL), and by GPT-4P (GPT-4 asked for the probability of a stereotype). It consists of error analyses in relation to the human annotations from a linguistic perspective, specifically, about the implicitness of the stereotypes. Below, we analyze the linguistic characteristics of the false negative and false positive texts encountered in all models, or only in FT-HL, FT-SL or GPT-4P.

All models There are 15 false negative (FN) instances and 12 false positive (FP) instances that failed to be correctly classified by all the models. These texts, indeed, have been annotated, respectively, as stereotypes and not stereotypes by all the annotators.

Regarding the **FN cases**, 13 out of 15 required conversational context for the message to be understood and the stereotype to be interpreted, whereas nine FN instances were annotated as implicit. Examples 2, 3 and 4 allude directly to immigrants, although their references have to be recovered from previous messages. The examples below also report the confidence obtained by FT-HL, the predicted soft label by FT-SL and the probability obtained with GPT-4P.

2. *Do you doubt it, those are votes for left-wing parties. Subsidy = vote.*
   [FT-HL: 0.93] [FT-SL: 0.04] [GPT-4P: 0.0]

3. *The ones for the dogs are not worth it?*
   [FT-HL: 0.71] [FT-SL: 0.19] [GPT-4P: 0.0]

4. *How strange* 🧐*, if they are little angels* 🥱
   [FT-HL: 0.81] [FT-SL: 0.12] [GPT-4P: 0.0]

The most recurrent form of implicit expression is through figures of speech, with seven cases. For instance, Example 2 suggests the instrumentalization of immigrants proposing a causal relation

between subsidies and votes, making a simile between these two elements. Another use of a figure of speech is found in Example 3, whose association between the hoax (the creation of a special beach for Muslim women) and a beach for dogs is expressed through a rhetorical question and could be interpreted as a joke aimed at dehumanizing the target group. Similarly, the use of irony in Example 4 is underpinned by a stereotyped perception of the target group, considering it is a response to a crime. Another case of implicit expressions of stereotypes is through the evaluation of the author's own situations (three cases), which entails a bias toward the target group, as seen in Example 8.

Finally, when observing the most repeated implicit expressions in the FN instances with the highest confidence scores (70% and higher), we found that irony is present in three out of six cases. 7 cases report confidence higher than 0.50, 13 cases have a predicted soft label lower than 0.40, and all cases report a GPT-4P probability lower than 0.20.

Regarding **FP instances**, a small number addressed the topic of immigration without conveying a stereotype, as in Example 5, when the author mentions the target groups, but no annotators perceived a stereotype.

5. *I know what the agreement says, but I don't know how many companies comply with it. If everyone is paid the same, whether they are seasonal workers, Moroccans, Huelva residents or sub-Saharans. That's what we want the labor inspection to find out.*
   [FT-HL: 0.67] [FT-SL: 0.67] [GPT-4P: 0.8]

Other recurrent topics are politics or opinions about other countries. For instance, in Example 6, which refers to the same hoax as Example 3, there is an attack to the politician who allegedly proposes the creation of beaches for Muslim women.

6. *He can take the Muslim women for a bath to his house in Galapagar*
   [FT-HL: 0.46] [FT-SL: 0.66] [GPT-4P: 1.0]

Four of the cases have a confidence score higher than 0.50, nine cases report a predicted soft label higher than 0.60, and all the cases report a GPT-4P probability higher than 0.80.

FT-HL The FN predictions present only five instances, with 100% annotator agreement, out of which four are implicit stereotypes and two need conversational context. Three of the implicit ones are news titles, in which the author relates an individual belonging to a specific social group and association to a negative event. There is a negative generalization of the actors of that specific event to all the members of the group, as seen in Example 7. The other implicit case corresponds to a rhyme implying that immigrants will expel the "natives". There are 73 cases of FPs.

7. *#URGENT: A foreign minor beats up a gay person in Pontevedra while shouting 'No to homosexuals.'*
   [FT-HL: 0.35]

FT-SL There are 29 tweets of FN instances, 25 of which with 100% annotator agreement, while there are 15 FP instances. Regarding FNs, 15 of them needed conversational context and only eight were implicitly expressed, seven of those needing also context. Out of the implicit stereotypes, all of them present metaphor (3), irony (2), a joke, the expression of a desire regarding immigrants underlying a stereotype and the case of a specific group of immigrants, which extrapolates indirectly the stereotype to the whole group. In comparison to FT-HL, FT-SL fails to predict more figures of speech, proportionally.

GPT-4 FN mispredictions comprise 35 instances, out of which 26 need conversational context and 21 are implicitly expressed, all of them with 100% annotator agreement. Out of the implicit stereotypes, 11 are presented in the form of figures of speech such as irony, metaphor, jokes and rhetorical questions. Six of them are evaluations of the author's feelings or desires, and four are imperative or 'call for action' expressions. Regarding the FP instances, there are 31 instances.

8. *I have the greatest respect for Muslims and Jews, I am the son of an emigrant and an immigrant myself. […] We cannot support anyone else, other than our poor families, while others get politically richer.*
   [GPT-4P: 0.2]

6. Discussion and Conclusion

In this paper, we performed a comparative analysis, both quantitative and qualitative, of the differences between the predictions made by various models and the annotations performed by humans. In particular, we wondered 1) about where the confidence of the FT-HL is lower and the probability of FT-SL and GPT-4 differ from the original soft label computed on the annotations; and 2) what are the characteristics of the texts in which the decisions of models and humans differ.

Regarding the first question, we demonstrated that the models in general align their predictions with humans, showing more confidence (FT-HL) and less variation in the distribution of the predicted probabilities (FT-SL and GTP-4P). It is worth remembering that in that cases all three annotators agree on the presence of stereotypes. In other

words, **models exhibit low confidence when annotators have more disagreement with each other**.

In turn, **there is more disagreement when the texts contain implicit stereotypes**, thus, the presence of implicit stereotypes contributes to the low confidence and less dense distribution of probabilities of the models. The models were better at predicting explicit stereotypes than implicit ones, even when accounting for different annotator perspectives.

Looking at the cases where humans do not agree with models predictions, we observed that **a majority of texts predicted as not-containing stereotypes with a higher confidence, include, instead, implicit stereotypes**. They are recurrently expressed through figures of speech (such as irony, metaphors) as well as through evaluations about the own author's feelings and thoughts about immigrants. Other cases in which the predictions differ from humans show **the need to refer to the conversational context to interpret the stereotype**.

On the other hand, among the instances predicted as positive with higher confidence but not annotated as stereotypes by humans (false positives), we encountered cases that addressed the topic of immigration without conveying a stereotype. This finding suggests that **the mere mention of the target group triggers its classification as stereotyped**.

Due to the high disagreement between humans, and also among models, when identifying implicit stereotypes, we propose, in future work, to establish specific categories that collect the different linguistic forms in which implicit stereotypes are expressed. In this way, we can operationalize implicitness in order to improve the annotation agreements and the recognition of stereotypes by models.

Moreover, since the lack of conversational contexts affects the identification of stereotypes, we propose to feed the models with the previous messages of the conversational thread, which is an input that humans had during the annotation process. In addition, since the explicit mention of a target group prompts the model to classify the instance as a stereotype, we propose to model different arguments or topics of stereotypes related to immigration as seen in Javier Sánchez-Junquera and Ponzetto (2021).

Finally, in this work we focused only on stereotypes related to immigrants, but we are curious to investigate what is the impact of implicit stereotypes on the perceptions of humans, as well as models, in the recognition of stereotypes towards other categories of people.

7. Acknowledgements

This work was supported by the international project STERHEOTYPES: STudying European Racial Hoaxes and sterEOTYPES funded by the Compagnia di San Paolo and VolksWagen Stiftung under the Challenges for Europe call (CUP: B99C20000640007); the SGR CLiC project (2021 SGR 00313) funded by the Generalitat de Catalunya, and the FairTransNLP-Language project (PID2021-124361OB-C33) funded by MICIU/AEI/10.13039/501100011033/ and by FEDER, UE.

Limitations

Although our approach wanted to consider the disagreement among annotators, we are aware that the number of annotators is low. However, we tried to guarantee a diversity in the process of annotation in accordance with Bender and Friedman (2018) and Cabitza et al. (2023). Annotators involved in the extension of the Spanish subset of the multilingual corpus (Bourgeade et al., 2023) were two linguistics students and a researcher on computational linguistics. They differed from country of provenience, gender and age.

8. Bibliographical References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Alejandro Ariza-Casabona, Wolfgang S. Schmeisser-Nieto, Montserrat Nofre, Mariona Taulé, Enrique Amigó, Berta Chulvi, and Paolo Rosso. 2022. Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural*, 69:217–228.

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics*, 35(4):495–503.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Camiel J. Beukeboom, Catrin Finkenauer, and Daniël H. J. Wigboldus. 2010. The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology*, 99(6):978–992.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang S. 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.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Dallas Card, Justin H Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1410–1420.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Andrea Cerase and Claudia Santoro. 2018. From racial hoaxes to media hypes: Fake news' real consequences. In *rom Media Hype to Twitter Storm: New Explosions and Their Impact on Issues, Crises, and Public Opinion*, p. vasterman edition, pages 333–54. P. Vasterman, Amsterdam.

Berta Chulvi, Lara Fontanella, Roberto Labadie, and Paolo Rosso. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*.

Katherine A. Collins and Richard Clément. 2012. Language and prejudice: direct and moderated effects. *Journal of Language and Social Psychology*, 31(4):376–396.

Francesca D'Errico and Marinella Paciello. 2018. Online moral disengagement and hostile emotions in discussions on hosting immigrants. *Internet Research*, 28(5):1313–1335.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4—27.

David L. (David Lewis) Hamilton. 1981. *Cognitive processes in stereotyping and intergroup behavior*. L. Erlbaum Associates.

Paolo Rosso Javier Sánchez-Junquera, Berta Chulvi and Simone Paolo Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8).

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On "scientific debt" in NLP: A case for more rigour in language model pre-training research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8554–8572, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

T. F. Pettigrew and R. W. Meertens. 1995. Subtle and blatant prejudice in western Europe. *European Journal of Social Psychology*, 25(1):57–75.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language.

Wolfgang S. Schmeisser-Nieto, Montserrat Nofre, and Mariona Taulé. 2022. Criteria for the annotation of implicit stereotypes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 753–762.

Mariona Taulé, Montserrat Nofre, Víctor Bargiela, and Xavier Bonet. 2024. Newscom-tox: a corpus of comments on news articles annotated for toxicity in spanish. *Language Resources and Evaluation*, pages 1–41.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.

9. Language Resource References

Ariza-Casabona, Alejandro and Schmeisser-Nieto, Wolfgang S. and Nofre, Montserrat and Taulé, Mariona and Amigó, Enrique and Chulvi, Berta and Rosso, Paolo. 2022. *DETESTS*. PID https://detestsiberlef.wixsite.com/detests/corpus.

Bourgeade, Tom and Cignarella, Alessandra Teresa and Frenda, Simona and Laurent, Mario and Schmeisser-Nieto, Wolfgang S. and Benamara, Farah and Bosco, Cristina and Moriceau, Véronique and Patti, Viviana and Taulé, Mariona. 2023. *A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads*.

Sánchez-Junquera, Juan Javier and Chulvi, Berta and Rosso, Paolo and Ponzetto, Simone Paolo. 2021. *Stereo Immigrants*. PID https://github.com/jjsjunquera/StereoImmigrants.

Mariona Taulé and Alejandro Ariza and Montserrat Nofre and Enrique Amigó and Paolo Rosso. 2021. *DETOXIS*. PID https://detoxisiberlef.wixsite.com/website/corpus.