Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 323-333 recibido 05-07-2024 revisado 16-07-2024 aceptado 29-07-2024

Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement

Resumen de la tarea de DETESTS-Dis en IberLEF 2024: DETEcción y clasificación de eSTereotipos raciales en eSpañol - Aprendizaje con desacuerdos

Wolfgang S. Schmeisser-Nieto[1,2], Pol Pastells[1], Simona Frenda[2,3], Alejandro Ariza-Casabona[1], Mireia Farrús[1], Paolo Rosso[4,5], Mariona Taulé[1]
[1]CLiC, UBICS, Universitat de Barcelona, Spain
[2]Computer Science Department, Universitá degli Studi di Torino, Italy
[3]aequa-tech, Turin, Italy
[4]PRHLT Research Center, Universitat Politècnica de València, Spain
[5]valgrAI- Valencian Graduate School Artificial Intelligence
{wolfgang.schmeisser, pol.pastells, alejandro.ariza14, mfarrus, mtaule}@ub.edu,
simona.frenda@unito.it, prosso@dsic.upv.es

Abstract: This paper presents an overview of the DETESTS-Dis shared task as part of the IberLEF 2024 Workshop on Iberian Languages Evaluation Forum, within the framework of the SEPLN 2024 conference. We proposed two hierarchical subtasks: In subtask 1, participants had to determine the presence of stereotypes in the texts. For subtask 2, participants had to decide which texts labeled with stereotypes were implicit stereotypes. The DETESTS-Dis dataset contains 12,111 comment sentences and tweets in response to newspaper articles and verified racial hoaxes involving immigration in Spanish. 15 teams signed up to participate, 6 of which sent runs, and 3 of them sent their working notes. In this paper, we provide information about the training and test datasets, the systems used by the participants, the evaluation metrics of the systems and their results.
Keywords: Stereotype Detection, Implicitness Detection, Immigration, Machine Learning.

Resumen: Este artículo presenta un resumen de la tarea DETESTS como parte del workshop IberLEF 2024, dentro del congreso SEPLN 2024. Propusimos dos subtareas jerárquicas: En la subtarea 1, los participantes tuvieron que determinar la presencia de estereotipos raciales en oraciones. En la subtarea 2, de las oraciones etiquetadas con estereotipo, los participantes tuvieron que decidir si este era implícito. El dataset DETESTS-Dis contiene 12.111 oraciones de comentarios y tuits respondiendo a artículos de periódicos y bulos raciales sobre inmigración en español. 15 equipos se registraron para participar, de los cuales 6 enviaron predicciones de sistemas y 3 de ellos enviaron artículos. En este artículo presentamos información sobre los datasets de entrenamiento y de prueba, los sistemas utilizados por los participantes, las métricas de evaluación y sus resultados.
Palabras clave: Detección de estereotipos, Detección de Implicitud, Inmigración, Aprendizaje Automático.

1 Introduction

Warning: *This paper contains examples of stereotypes that may be offensive to some readers.*

In this paper, we introduce the second edition of the DETESTS task, first presented at IberLEF 2022 (Ariza-Casabona et al., 2022). The aim of the new edition, DETESTS-Dis[1] (DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement), is to detect the presence of stereotypes and classify them as either explicitly or implicitly manifested in social media mes-

[1]https://detests-dis.github.io

ISSN 1135-5948 DOI 10.26342/2024-73-24

sages and comments on news articles. The primary novelty of this task lies in the release of disaggregated annotations within the dataset, aimed at encouraging participants to develop models that account for potential disagreements among annotators. Recognizing stereotypes is inherently subjective and can lead to varying opinions due to annotators' sensitivity, cultural background, age, and gender. Therefore, the decision regarding which data to train the models with is non-trivial, as research has shown that variability in agreement affects the confidence of the models in making decisions (Schmeisser-Nieto et al., 2024). Therefore, participants in the task were given all the available annotations, that is, the gold standard (hard labels), as well as the pre-aggregated labels, following the Learning with Disagreement (Uma et al., 2021; Leonardelli et al., 2023) and the Perspectivism paradigm (Cabitza, Campagner, and Basile, 2023). The texts consist of tweets (now known as X posts) published in response to verified racial hoaxes (a type of fake news in which there is a deliberate intention to harm an individual or a group based on their origin, religion, or ethnicity), and sentences extracted from comments on online news articles related to immigration. All texts are in Spanish.

The DETESTS-Dis task took place as part of IberLEF 2024, the 6th Workshop on Iberian Languages Evaluation Forum at the SEPLN 2024 conference (Chiruzzo, Jiménez-Zafra, and Rangel, 2024).

2 Background

One of the main components that reinforces toxic and hate speech are stereotypes. Understanding how they emerge and spread is crucial for tackling this issue, since stereotypes are not always expressed explicitly. The presence of stereotypes on social media and the need to identify and mitigate them is leading to the development of systems for their automatic detection, especially in news comments and tweets. Consequently, this emerging task is garnering increasing interest from the NLP community. For instance, in the first edition of DETESTS, 39 teams submitted runs, five of which also submitted working papers.

In social psychology, a stereotype is a set of beliefs about others perceived as belonging to a different social category. The stereotype oversimplifies the group and generalizes a characteristic, applying it to all its members (Allport, Clark, and Pettigrew, 1954). Stereotypes are a cognitive component and, as with prejudice, their emotional counterpart, they model behavior toward others (Fiske, 1998). Stereotypes are expressed in language through several communication acts, which can be explicit, that is, transparent and manifest (see Example (4)), or implicit, when a process of inference is necessary for the stereotype to be perceived (see Example (3)) (Schmeisser-Nieto, Nofre, and Taulé, 2022).

Numerous works in languages such as English, Italian, French, Spanish and Dutch have focused on stereotype recognition against vulnerable social groups, e.g., women and immigrants. For instance, Automatic Misogyny Identification (Fersini, Nozza, and Rosso, 2018) presents a classification subtask in which one of the categories of misogyny is Stereotype and Objectification understood as a fixed and oversimplified image or idea of a woman. EXIST (Rodríguez-Sánchez et al., 2021; Rodríguez-Sánchez et al., 2022; Plaza et al., 2023) tackled the topic of sexism in social networks, while more specifically, studies on the detection of gender stereotypes have also been conducted (Cryan et al., 2020; Chiril, Benamara, and Moriceau, 2021). Among the approaches to identifying stereotypes within narratives, there are studies focusing on microportraits, in which a description of Muslim people is provided in a single text (Fokkens et al., 2018). Sap et al. (2020) addresses the issue of social bias frames driven by stereotypes. Evalita 2020's HaSpeeDe 2 task included a subtask on the identification of immigrants, Muslims and Roma (Sanguinetti et al., 2020). Narrowing down on the topic of immigration, Sánchez-Junquera et al. (2021) put forward a classification of stereotypes manifested in political debates. DETESTS-Dis is therefore an innovative proposal that goes a step further in stereotype identification by incorporating the identification of implicitly expressed stereotypes.

Finally, in recent years, various evaluation tasks have incorporated a learning-with-disagreement approach, providing participants with disaggregated annotations (Uma et al., 2021; Plaza et al., 2023; Leonardelli et al., 2023).

3 Task Description

This task is designed hierarchically by chaining two binary-classification subtasks:

Subtask 1: Stereotype Detection
The first subtask aims to determine whether a comment or sentence contains any stereotype, considering the full distribution of labels provided by the annotators. This subtask follows the SemEval 2021 Task 12 (Uma et al., 2021) proposal about learning with disagreement, in which the authors state that there is not necessarily a single gold label for every sample in the dataset. This fact is particularly evident when multiple contradictory annotations arise at the data labeling stage due to "debatable, subjective, or linguistic ambiguity". The actual gold label of this subtask is left as a proxy to determine the subset of comments that will be evaluated in the subsequent subtask. Example (1) shows an instance with a stereotype and (2) without one.

(1) [...] Los inmigrantes ilegales tienen más derechos que los españoles y estamos HARTOS!
'*[...] Illegal immigrants have more rights than Spaniards and we are FED UP!.*'

(2) De echo todos los países de mayoría musulmana tienen la sharia como fuente de derecho en mayor o menor grado, todos son teocráticos.
'*In fact, all Muslim-majority countries have Sharia as a source of law to a greater or lesser degree, they are all theocratic.*'

Subtask 2: Implicitness Identification

This optional subtask introduces a hierarchical binary classification problem to identify whether stereotypes in the text are explicit or implicit. Implicit stereotypes require inference, as they are not directly expressed. These can manifest through various communicative strategies such as metaphor, irony, and other figures of speech, evaluations of the in-group, or overgeneralizations based on some members' features. For instance, an implicit stereotype is illustrated in (3) through an in-group evaluation, while (4) provides an example of an explicit stereotype.

(3) Acabaremos siendo una minoría en nuestro propio país.
'*We will end up being a minority in our own country.*'

(4) Como no les vota ni el Tato, pretenden pillar el voto de los inmigrantes ilegales.
'*Since no one votes for them, they want to get the vote of illegal immigrants.*'

4 Dataset

The DETESTS-Dis dataset consists of two text genres from two different corpora: comments on news articles (DETESTS corpus along with an extension consisting of a newly extracted and annotated set) and posts on Twitter reacting to hoaxes (StereoHoax-ES corpus) about immigrants.

The DETESTS dataset comprises the released corpus for the DETESTS task at IberLEF 2022 (Ariza-Casabona et al., 2022) and an extension created specially for this task. It is therefore made up of three parts: one part from the NewsCom-TOX corpus (Taulé et al., 2024), with 3,306 sentences, another part from the StereoCom corpus, with 2,323 sentences, which was created following the same methodology as NewsCom-TOX, and finally, 1,133 additional sentences were specifically extracted and annotated for the test set of this task, following the methodology used in NewsCom-Tox and StereoCom. All parts consist of comments published in response to different articles extracted from Spanish online newspapers (ABC, elDiario.es, El Mundo, NIUS, etc.) and discussion forums (such as Menéame). In the case of NewsCom-TOX, the comments were extracted from news articles published from August 2017 to August 2020, while in the case of StereoCom, the extraction period ranges from June 2020 to November 2021. For both corpora, DETESTS and NewsCom-TOX, the articles were manually selected considering their controversial subjects, their potential toxicity, and the number of published comments (minimum of 50). A keyword-based approach was used to search for articles mainly related to racism and xenophobia. Since the NewsCom-TOX corpus was designed primarily to study toxicity and not stereotypes, we used only the part of the corpus with the highest percentage of stereotypes. The comments were selected in temporal order as they appear in the conversational thread. Each comment was segmented by punctuation into sentences, and the comment to which every sentence belongs and its position within the comment are indicated.

The StereoHoax-ES dataset was created within the framework of the STERHEO-

TYPES project (Bourgeade et al., 2023), which brought together international research units based in Italy, France, and Spain. The corpus used for the second edition of DETESTS is the Spanish subset of the StereoHoax multilingual dataset.

It contains tweets, retrieved from Twitter in 2021, reacting to hoaxes published online that aimed to disseminate false news against immigrants in Spain. These tweets were collected taking into account their conversational thread. For the Spanish subset, 5,349 tweets from 449 conversational heads, i.e., the root nodes, were retrieved.

The collection of these tweets started with the manual identification of 72 anti-immigrant hoaxes on debunking websites, such as Maldita.es[2] and Newtral[3]. Using the titles, keywords, and content of the hoaxes, they were searched using the Twitter API v2 for Academia, collecting conversations related to them. The conversational thread is represented by a conversational head (the tweet starting the conversation), direct replies and replies to replies.

	DETESTS	**StereoHoax-ES**
Text	Sentence	Tweet
Level 1	Previous Sentences	–
Level 2	Parent Comment	Parent Tweet
Level 3	Root Comment	Root Tweet
Level 4	News Title	Hoax

Table 1: Context levels for DETESTS and StereoHoax-ES.

Both corpora, DETESTS and StereoHoax-ES, have a thread structure. The first comment or tweet is the root of the thread. The root can then have multiple responses, forming a tree structure. We identified a range of contexts to which annotators had access to provide them to the models. We structured the contexts into four levels, summarized in Table 1:

1. Previous sentences in the same comment (level 1). This level is only available for DETESTS, as StereoHoax-ES tweets were not split into sentences. Additionally, this level does not apply to the first sentence of each comment, which constitutes 45% of sentences in the DETESTS.
2. Previous text in the thread (level 2). This level is absent for the first comment in each thread, accounting for 45% of comments in DETESTS and 8% of tweets in StereoHoax-ES.
3. Root text (level 3). This level does not exist for the first comment of each thread and is identical to the previous comment for the second comment on each thread. It is missing in 45% of the comments and 16% of the tweets. Note that DETESTS has full threads, so the comments missing level 2 and the ones missing level 3 are the same, while for StereoHoax-ES they are different, although overlapping, sets.
4. News title for DETESTS or fake news text for StereoHoax-ES (level 4). This level is always present and differs from the others in that it does not represent an instance of the dataset, but an external reference.

4.1 Annotation Scheme

Both datasets were fully annotated with the label *stereotype* on the presence or absence of at least one stereotype. When the annotators decided on the positive class, they had to decide on the *implicitness* of the stereotype. The binary values were 0 for the absence of the feature and 1 for its presence. Therefore, an implicit stereotype was annotated with 1. Annotators also had access to the conversational context. The instructions provided in the annotation guidelines defined each label as:

Stereotype: A stereotype is a cognitive mechanism consisting of a set of beliefs regarding another social group, namely the outgroup, which is perceived as different. Forming these beliefs involves a homogenization process, where one characteristic of an individual or part of the group is generalized to the entire group. This generalization typically occurs based on factors such as place of origin, ethnicity, or religion. Stereotypes can be expressed explicitly or implicitly.

Implicitness: An implicit stereotype refers to a stereotype that is not transparent; it is present, but not evidently so. In an implicit message, part of the meaning is not fully expressed. There is a process of inference undertaken by the reader to interpret the stereotype. Implicit stereotypes can manifest through various

[2]https://maldita.es/
[3]https://www.newtral.es/

linguistic strategies such as metaphors, irony, humor, entailments, evaluations of the in-group as a consequence of the out-group, and generalizations. In contrast, explicit stereotypes are expressed with precision, detail and clarity, leaving no room for doubt or confusion. Explicit stereotypes are normally copulative sentences with adjectives and predicative using the habitual aspect, as in Examples (5) and (6), respectively (Schmeisser-Nieto, Nofre, and Taulé, 2022).

(5) Immigrants are thieves.

(6) Immigrants do the jobs we don't want to do.

4.2 Annotation Process

Each instance was annotated in parallel by three annotators, consisting of a researcher in linguistics and two linguistics students trained for the task. Weekly meetings were held to discuss doubts and controversial cases with the involvement of a senior researcher.

For the inter-annotator agreement test, a Fleiss' kappa coefficient was calculated in both datasets. For the DETESTS dataset, there is a moderate inter-annotator agreement of 0.57 on the presence of stereotypes and of 0.41 for the implicit forms. For the StereoHoax-ES dataset, a substantial agreement of 0.75 was reached on the presence of stereotypes, whereas a slight agreement of 0.15 was obtained on implicitness.

Majority voting was used to determine the hard labels. Thus, at least two annotators needed to label a text as containing at least a stereotype to be labeled as such. For this task, both datasets were released in their disaggregated forms to provide participants with the opportunity to conduct experiments while considering disagreements among the annotators. Furthermore, an example of a soft label was given to the participants, computed as the softmax normalization of the three annotators (Uma et al., 2020).

Table 2 presents the label distribution across both corpora. The texts annotated as containing stereotypes by 0 or 1 annotators conform to the "No Stereotype" class after majority voting. Otherwise, the majority vote leads to a "Stereotype" label.

4.3 Training and Test Set

We provided participants with 82% of the DETESTS-Dis dataset to train and validate their models (5,629 sentences and 4,277 tweets), and the remaining 18% (1,133 sentences and 1,072 tweets) was used as a test set to evaluate their performance[4]. The subsets were stratified to maintain the same distribution of implicit and explicit stereotypes. To avoid data leakage, we separated tweets extracted from different hoaxes into different sets and used the newly annotated sentences for the test set.

	DETESTS	**StereoHoax-ES**
Total	6,762	5,349
No Stereotype	4,840	3,745
0 votes	3,963	3,249
1 vote	877	496
Stereotype	1,922	1,604
2 votes	178	359
3 votes	1744	1245
Implicit	1,556	344

Table 2: Label distribution for DETESTS and StereoHoax-ES.

The training set consisted of the following columns:

- The source for the text: *detests* or *stereohoax*.
- A unique identifier.
- A comment identifier, to group the DETESTS sentences from the same comment.
- The text to be classified, be it a sentence or a tweet.
- The four levels of context: level1, a pointer to the previous sentence id; level2, the previous tweet or comment, with comment id as a pointer; level3, the first tweet or comment, referring to comment id; level4, a pointer to the news text or the racial hoax identifier, provided in a different file.
- The three individual annotations for the presence of a stereotype and implicitness, as well as the majority voting (hard label) and the softmax normalization (soft label) for both.

The test set given to the participants had the source, id, comment id, text and context levels.

Given the restrictions posed by EU GDPR and to avoid any conflict with the sources

[4]The dataset is available upon request.

of the comments regarding their intellectual property rights (IPR), both training and test data were made available for academic purposes only, and participants therefore accessed the data with a password by filling in an online form[5]. No user data was disclosed, since all the data were anonymized by removing all personal information such as @user and generating new IDs for the texts coming from Twitter.

5 Evaluation metrics and baselines

Subtask 1 was evaluated as in the LeWiDi shared tasks at SemEval 2021 and 2023 about learning with disagreement (Uma et al., 2021; Leonardelli et al., 2023). First, the models that output hard labels were compared to the gold standard using the binary F1 metric. The second evaluation metric was the cross-entropy between the system soft label values and the soft labels generated from the average votes of the annotators.

Subtask 2 was a binary hierarchical classification problem. We used the ICM metric (Amigo and Delgado, 2022), an information-theoretic-based metric that considers both the hierarchical structure and the class specificity. It applies to both hard and soft labels. The ICM metric was the official metric considered for the ranking for both hard labels (ICM) and soft labels (ICM-Soft). The implementation of the official metrics (F1, cross-entropy and ICM) was based on PyEvALL (Amigó et al., 2017).

5.1 Baselines

For both subtasks 1 and 2 with hard labels, we used six baselines: AllOnes, AllZeros, RandomClassifier, TFIDF+SVC, FastText+SVC and BETO. For simplicity, in the soft labels option, we only show results for the top-performing and well-established baseline, i.e. BETO. Subtask 2 baselines were a hierarchical extension from the ones of subtask 1. The sentences predicted to contain a stereotype in the first subtask were used to infer the implicitness of said stereotype. The baselines were as follows:

AllOnes: A non-informative baseline that classified all instances as the positive class.

AllZeros: Analogous to AllOnes, it maps all instances to the negative class.

RandomClassifier: A weighted random classifier with probabilities based on the train set label distribution.

TFIDF+SVC: A Term Frequency-Inverse Document Frequency (TFIDF) vectorizer was used to extract features based on the 10,000 unigrams (lowercased in Unicode). A Support Vector Classifier (SVC) with a linear kernel was used to determine the predictions.

FastText+SVC: A word vector extractor based on the FastText algorithm followed by a mean pooling operation for sentence-level representation, followed by a SVC.

BETO: A Spanish BERT model[6] (Cañete et al., 2020) fine-tuned for a classification task—hard labels—and a regression task—soft labels.

All baselines were implemented in the Python language, and the code was accessible to the participants from the task GitHub repository[5].

6 Systems Overview

The DETESTS-Dis shared task received submissions from six teams for subtask 1 with hard labels and three teams with soft labels, and from four teams for subtask 2 with hard labels and three teams with soft labels. Participants were allowed to provide up to three submissions per subtask and type of label (hard or soft). Table 3 shows a summary of the number of teams and runs for each task. The proposed systems are summarized below.

Task	Teams	Runs
1 - Hard Labels	6	15
1 - Soft Labels	3	7
2 - Hard Labels	4	8
2 - Soft Labels	3	5

Table 3: Number of teams and runs for each task.

The top scoring team for subtask 1 with hard labels, **Brigada Lenguaje** used the embeddings from a fine-tuned BETO model with a linear classifier for each annotator. The loss function was the sum of the three loss functions for each annotator. The hard labels prediction was obtained by a majority voting, and the soft labels one with a softmax normalization of the predictions for each annotator.

[5]https://github.com/clic-ub/DETESTS-Dis

[6]dccuchile/bert-base-spanish-wwm-cased

UC3M-SAS (González, García-Chicangana, and Galvis, 2024) fine-tuned the RoBERTa-base-bne[7] model (Fandiño et al., 2022), placing first for subtask 1 with soft labels, and the BETO model, third for the same subtask. UC3M-SAS used data augmentation just in the training set. They used back-translation from Spanish to English separately in the tasks with hard or soft labels. For the hard labels tasks, they used back-translation with the texts containing stereotypes, as well as synonym substitution, while with soft labels they augmented the texts with 1 or 2 annotators labeling them as containing stereotypes, given that the cases with full agreement were more common.

The **EUA** team did the best in subtask 2, with both hard and soft labels. They also used back-translation, from Spanish to English, and considered the previous sentence for DETESTS (level 1) and the first tweet for StereoHoax-ES (level 3). The winning strategy with hard labels was to use a regression task to predict soft labels and map them to 0 or 1, while the best approach for soft labels was a multitask approach, with one prediction for each annotator.

I2C-Huelva (Carrejón-Naranjo et al., 2024) obtained the second and third positions in task 1 by fine-tuning the RoBERTA-base-bne and BETO models respectively. They fine-tuned a different model for each corpus and annotator, and used majority voting to determine the ensemble output. They also used back-translation, with a four step approach, translating first from Spanish to English and from English to German and back from German to English and from English to Spanish.

TaiDepZai999_UIT_AIC (Tai, 2024) did an ensemble with BETO, XLM-RoBERTa (Conneau et al., 2020) and twitter-roberta-base-hate[8] (Barbieri et al., 2020). They also included the contexts by prepending the first two levels to the text and appending level 3 at the end (see Table 1 for information about each level).

VINE Bias Busters sent three runs for subtask 1 with hard labels. Their best run was a fine-tuning of BETO, followed by a fine-tuning of RoBERTa, and a prompting of GPT-3.5-Turbo.

[7] PlanTL-GOB-ES/roberta-base-bne

[8] cardiffnlp/twitter-roberta-base-hate

7 *Systems Results*

7.1 Subtask 1

Rank	Team	F1
1	Brigada Lenguaje I	0.724
2	I2C-Huelva I	0.712
4	EUA II	0.691
	BETO	0.663
7	UC3M-SAS II	0.641
8	TaiDepZai999_UIT_AIC I	0.630
	AllOnes	0.589
12	VINE Bias Busters I	0.581
	TFIDF+SVC	0.297
	RandomClassifier	0.297
	FastText+SVC	0.297
	AllZeros	0.000

Table 4: Evaluation results in subtask 1 with hard labels. Roman numerals show the run number for each team.

Rank	Team	Cross Entropy
1	UC3M-SAS I	0.841
2	EUA II	0.850
	BETO	0.893
4	Brigada Lenguaje I	0.938

Table 5: Evaluation results in subtask 1 with soft labels.

Tables 4 and 5 show the ranking of the teams participating in subtask 1 with hard and soft labels. Five runs got a better result than the BETO baseline with hard labels and three with soft labels. The entropy of the gold standard, and the minimum possible cross entropy, was 0.255. In general, for both metrics, the variance of the scores among all the participants is low, and in particular, the cross entropy values are high.

Looking specifically at the performance in the hard label context, we report in Figure 1 the confusion matrix for the best run of the first four teams, along with the Gold Standard and the BETO baseline. The first team achieved a higher binary F1 by having a lower number of False Positives (FP) and a higher number of True Negatives (TN) than the second team, who reported a higher sensibility to the recognition of positive class.

7.2 Subtask 2

Tables 6 and 7 show the team ranking for subtask 2 with hard and soft labels. The maximum possible ICM value, the one achieved with the Gold Standard itself, was 1.380; and

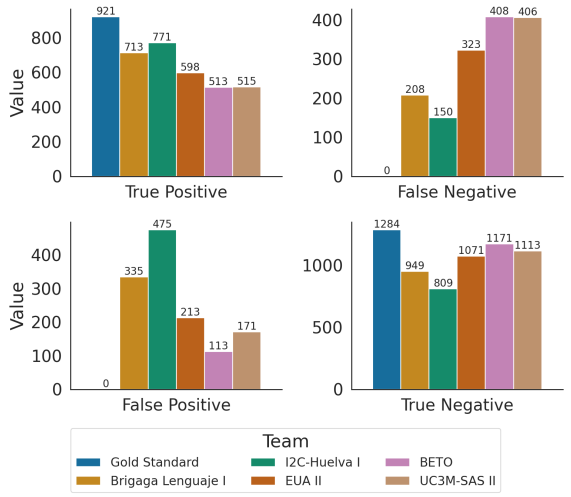

Figure 1: Confusion matrix for the best run of the first four teams in subtask 1 with hard labels, together with the Gold Standard and the BETO baseline.

Rank	Team	ICM	ICM Norm
	BETO	0.126	0.546
1	EUA II	0.065	0.524
4	Brigada Lenguaje I	−0.240	0.413
	TFIDF+SVC	−0.275	0.400
	FastText+SVC	−0.412	0.351
	AllZeros	−0.797	0.211
	RandomClassifier	−1.056	0.117
	AllOnes	−1.210	0.061

Table 6: Evaluation results in subtask 2 with hard labels.

the maximum possible ICM-Soft was 4.651. Both values correspond to a normalized metric of 1.000. No team beat the BETO baseline with hard labels and a single team (EUA) got three runs on top of the BETO baseline with soft labels. Figure 2 shows the confusion matrix for subtask 2 with hard labels. Both the BETO baseline and EUA's best run show a significant number of implicit stereotypes classified as not containing a stereotype. Given the nature of the hierarchical subtask, the False Negatives (FN) of the first class do not have any chance to be classified in the second subtask. This leads to a low recall of the implicit stereotypes. Therefore, subdividing the problem into two hierarchical tasks, where the first is to detect a stereotype and the second to classify it as implicit or explicit, does not seem to be the right approach. Tackling the problem as a flattened multi-class classification task could be an interesting approach for future work.

8 *Conclusion and Future Work*

The DETESTS-Dis task comprised two hierarchical binary-classification subtasks. Subtask 1 involved detecting the presence of stereotypes in social media texts, while subtask 2 consisted of deciding whether the stereotypes in those texts were implicit or explicit. Apart from the texts, the instances also included different levels of the conversational thread. The participants were given the gold standard annotations (hard labels), as well as the disaggregated annotations (soft labels). A total of 15 teams signed up to participate, of which six sent runs and three sent a working paper.

Various teams employed data augmentation techniques to enrich the positive class, e.g., back-translation, which seems to be beneficial for their systems. Despite the availability of different contextual levels, only a few teams employed them, and those that did, showed no significant improvement in terms of performance. Even when most teams did not submit runs for the soft label tasks, the proposed models were designed to exploit all the available annotations. This was an important goal in our shared task, since we wanted to encourage the community to implement inclusive models that were aware of disagreement regarding subjective phenomena like stereotype detection.

Looking at the performance of the models, we notice a tendency to misclassify the presence of stereotypes (subtask 1), except for the I2C-Huelva's model, which appears more sensitive to the positive class. In subtask 2, the models tend to identify better the cases of explicit stereotypes, except for the EUA's model.

Finally, none of the proposed approaches were computationally intensive. Therefore, the use of large language models with higher computational capacity and proper prompting techniques could improve the results. Considering the outcome of this task, it can be concluded that the detection of implicit stereotypes remains a significant challenge. Consequently, future research should focus on exploring how specific techniques, such as leveraging context, data augmentation, or undersampling, could enhance the detection of implicit stereotypes, particularly in addressing data imbalance.

Rank	Team	ICM-Soft	ICM-Soft Norm
1	EUA III	−0.900	0.403
	BETO	−1.124	0.379
4	UC3M-SAS II	−1.250	0.366
5	Brigada Lenguaje I	−1.684	0.319

Table 7: Evaluation results in subtask 2 with soft labels.

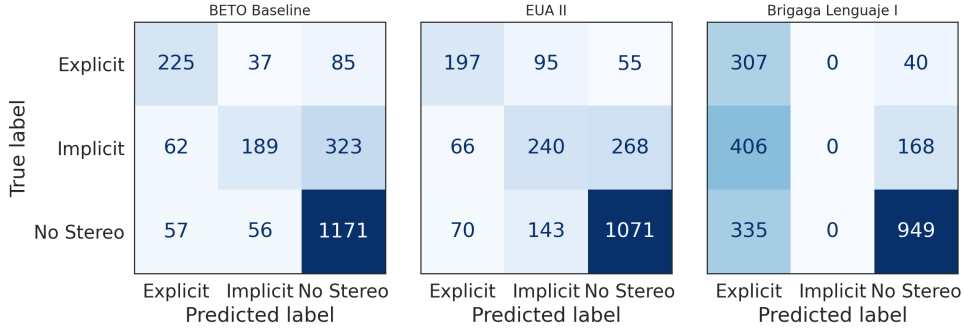

Figure 2: Confusion matrices for subtask 2 with hard labels best three teams.

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.

References

Allport, G. W., K. Clark, and T. Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Amigó, E., J. Carrillo-de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodríguez-Vidal, and F. Verdejo. 2017. Evall: Open access evaluation for information access systems. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1301–1304.

Amigo, E. and A. Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819.

Ariza-Casabona, A., W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, and P. Rosso. 2022. Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish. *Procesamiento del Lenguaje Natural*, 69:217–228.

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November. Association for Computational Linguistics.

Bourgeade, T., A. T. Cignarella, S. Frenda, M. Laurent, W. S. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, and M. Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*. In press.

Cabitza, F., A. Campagner, and V. Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Carrejón-Naranjo, M., M. Guerrero-García, J. Mata-Vázquez, and V. Pachón-Álvarez. 2024. I2C-Huelva at IberLEF-2024 DETESTS: Learning from Divergence to Identify Explicit and Implicit Racial Stereotypes in Spanish Texts. In *IberLEF@ SEPLN*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chiril, P., F. Benamara, and V. Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Cryan, J., S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–11.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Fersini, E., D. Nozza, and P. Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Fiske, S. 1998. Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.). *The handbook of social psychology*, pages pages 357–411.

Fokkens, A., N. Ruigrok, C. Beukeboom, G. Sarah, and W. Van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3734–3741.

González, A., D. S. García-Chicangana, and S. R. Galvis. 2024. UC3M-sas at IberLEF2024 DETESTS-Dis tasks. In *IberLEF@ SEPLN*.

Leonardelli, E., G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, and M. Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada, July. Association for Computational Linguistics.

Plaza, L., J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. 2023. Overview of exist 2023–learning with disagreement for sexism identification and characterization. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 316–342. Springer.

Rodríguez-Sánchez, F., J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of EXIST 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Rodríguez-Sánchez, F., J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, and P. Rosso. 2022. Overview of EXIST 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69(0):229–240.

Sánchez-Junquera, J. J., B. Chulvi, P. Rosso, and S. P. Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying

stereotypes about immigrants. *Applied Sciences*, 11(8).

Sanguinetti, M., G. Comandini, E. di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In V. Basile, D. Croce, M. Di Maro, and L. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765. CEUR Workshop Proceedings (CEUR-WS.org). Conference date: 17-12-2020.

Sap, M., S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July. Association for Computational Linguistics.

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

Schmeisser-Nieto, W. S., P. Pastells, S. Frenda, and M. Taule. 2024. Human vs. machine perceptions on immigration stereotypes. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8453–8463, Torino, Italia, May. ELRA and ICCL.

Tai, L. D. 2024. TaiDepZai999_UIT_AIC at IberLEF 2024 DETEST-Dis task: Classification of racial stereotypes in Spanish With Ensemble Learning Methods and BERT-based Adapter Head. In *IberLEF@ SEPLN*.

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

Uma, A., T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, and M. Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online, August. Association for Computational Linguistics.

Uma, A., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177, Oct.