I2C-Huelva at IberLEF-2024 DETESTS-Dis: Learning from Divergence to Identify Explicit and Implicit Racial Stereotypes in Spanish Texts

Manuel Cerrejón-Naranjo[1,*], Manuel Guerrero-García[1], Jacinto Mata-Vázquez[1] and Victoria Pachón-Álvarez[1]

[1]*I2C Research Group ,University of Huelva, Spain*

Abstract

This paper presents the approaches developed for detecting and identifying racial stereotypes in Spanish texts using advanced Natural Language Processing (NLP) and Deep Learning techniques, incorporating Learning with Disagreement for enhanced robustness. The major contribution of this work is the demonstration of the effectiveness of transformer-based ensemble classifiers to recognize both explicit and implicit stereotypes. By leveraging the strengths of multiple models, the proposed method achieves better performance than using a single model alone. Additionally, the importance of selecting appropriate hyperparameters during the model training process was highlighted by the results. Through rigorous experimentation and evaluation, optimal hyperparameter combination where identified.

In our experiments, we utilized a preprocessed and annotated corpus of Spanish texts and applied data augmentation techniques, such as back-translation, to balance the dataset. Furthermore, we incorporated the "Learning With Disagreement" (LeWiDi) approach, which uses the discrepancies between different models to improve the classification system. The results obtained demonstrate significant improvements in F1-Score, underscoring the potential application of these methods in moderating content on social media and other digital platforms. With this strategy, we achieved second place in Task 1 using an ensemble consisting of 3 models, one for each annotator, based on RoBERTa. In Task 2, we reached the seventh position, using the same approach.

Keywords

Learning With Disagreement Stereotypes, Natural Language Processing, Deep Learning, Transformers, Data Augmentation, Hyperparameter Optimization, Ensemble,

1. Introduction

Racial stereotypes are oversimplified and generalized beliefs about individuals based on their perceived social group membership [1]. In contemporary society, social media platforms have become a prominent venue for expressing opinions, including those related to immigration, a topic of significant public interest. However, the anonymity and reach of these platforms have also facilitated the proliferation of toxic and stereotypical comments. The ability to detect and classify such stereotypes, whether explicit or implicit, is crucial for mitigating bias and promoting healthier online discourse.

IberLEF 2024, September 2024, Valladolid, Spain

manuel.cerrejon886@alu.uhu.es (M. Cerrejón-Naranjo); manuel.guerrero790@alu.uhu.es (M. Guerrero-García); mata@uhu.es (J. Mata-Vázquez); vpachon@dti.uhu.es (V. Pachón-Álvarez)

In this paper, we present our systems developed for the DETESTS-Dis (DETEction and classification of racial STereotypes in Spanish - Learning with Disagreement) shared task at IberLEF 2024 [2] [3], held at the SEPLN 2024 conference. The goal of this task is to detect and classify both explicit and implicit stereotypes in Spanish texts from social media and news article comments. This edition incorporates the Learning with Disagreement paradigm by providing both hard and pre-aggregated labels to manage annotator disagreements.

Given the recent advancements in pre-trained language models for text classification, our approach centers on fine-tuning these models to adapt them to the task of stereotype detection. Additionally, we address the challenge of class imbalance using various data balancing techniques. By leveraging ensemble methods, we aim to improve the F1-Score of stereotype detection in our models.

The paper is structured as follows: Section 2 reviews related works in stereotype detection and classification. Section 3 describes the dataset provided by the task organizers and outlines its use in our experiments. Section 4 details our methodological approach and experimental setup, followed by the results obtained. Section 5 discusses the submitted runs and evaluation findings. Finally, Section 6 presents our conclusions and potential future directions for research.

2. Related works

Numerous studies have targeted stereotype detection and classification within specific social groups, including women and immigrants. For instance, Fersini et al. [4] introduced the Automatic Misogyny Identification task, which included a subtask on stereotype and objectification of women. Similarly, Rodríguez-Sánchez et al. [5] developed the EXIST dataset to address sexism in social networks, exploring machine learning and deep learning techniques for automatic detection of sexist expressions and attitudes on Twitter.

Chiril [6] and Cryan et al. [7] further investigated the detection of gender stereotypes, contributing valuable datasets and methodologies. Fokkens et al. [8] studied microportraits to identify stereotypes in narratives about Muslim individuals, while Sap et al. [9] examined social bias frames driven by stereotypes. Sanguinetti et al. [10] expanded this work to include stereotypes about immigrants, Muslims, and Roma in the HaSpeeDe 2 task.

Specifically, Sánchez-Junquera et al. [11] focused on the classification of stereotypes related to immigration within political debates, highlighting the complexity of implicitly expressed stereotypes. The first DETESTS task [12] made significant strides in detecting and classifying stereotypes in Spanish texts, setting a precedent for the current DETESTS-Dis task. This task aims to further advance the field by incorporating learning with disagreement techniques to better handle annotator disagreement and enhance detection accuracy.

Moreover, Leonardelli et al. [13] explored the potential of integrating disagreement learning to improve stereotype detection systems, providing a framework that is particularly relevant for the DETESTS-Dis task. Uma et al. [14] also contributed to this paradigm, demonstrating how leveraging annotator disagreement can lead to more robust and reliable classification models.

These previous works underscore the importance and challenge of detecting and classifying stereotypes in various contexts. The DETESTS-Dis task builds upon these foundations, aiming to address both explicit and implicit stereotypes in social media and news comments, thereby

contributing to the growing body of research in this critical area.

3. Dataset Description and Task Objectives

The training dataset provided by the organizers contains 9906 comments published in response to two different sources, Detests and Stereohoax. There are 18 information attributes available, which are: source, id, comment-id, text, level1, level2, level3, level4, stereotype-a1, stereotype-a2, stereotype-a3, stereotype, stereotype-soft, implicit-a1, implicit-a2, implicit-a3, implicit, and implicit-soft. The dataset was split while maintaining the class stratification into training (70%), valid (20%) and test (10%).

- The objective of Subtask 1 is to determine if the sentences in a comment contain at least one stereotype or none, considering the complete distribution of provided labels.

Table 1
Example instances of Subtask 1

id	text	st_a1	st_a2	st_a3	st
s_77	Seguidamente va a reventar. Apretando mucho las tuercas	0	0	0	0
d_684_01	600000 ilegales y no pasa nada	1	1	1	1

Table 2
Class Distribution of Subtask 1

Class	Train Dataset	Valid Dataset	Test Dataset
0	5110	1461	730
1	1823	521	261
Total	6933	1982	991

For the second task, the dataset contains 2605 comments, with the same labels than the first task

- For the Subtask 2, the objective is to determine whether the stereotype in a comment is expressed explicitly or implicitly, considering the complete distribution of provided labels.

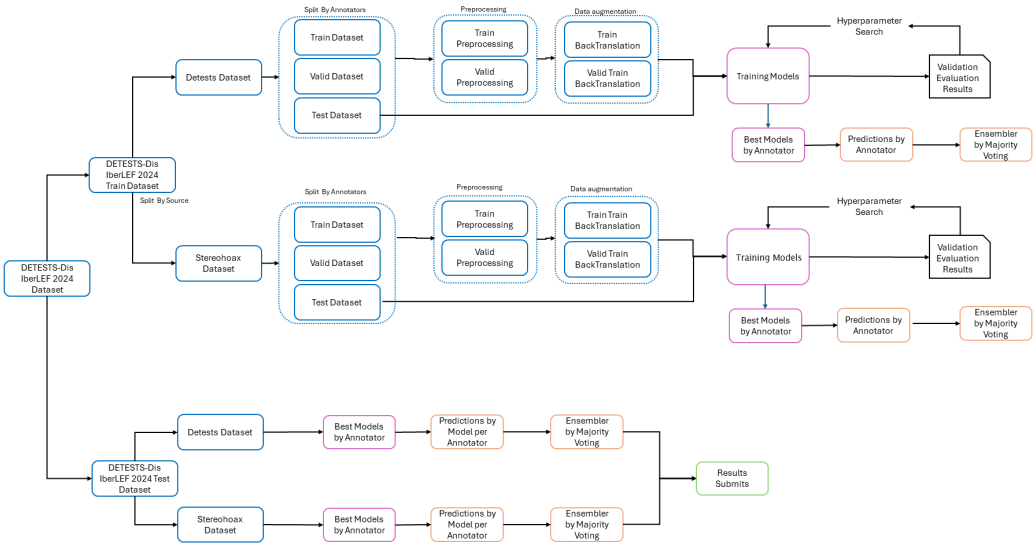
Table 3
Example instances of Subtask 2

id	text	im_a1	im_a2	im_a3	im
s_3315	Joder... Ahora todo áfrica aquí?? Que asco	1	1	0	1
d_684_01	600000 ilegales y no pasa nada	0	0	0	0

Table 4
Class Distribution of Subtask 2

Class	Train Dataset	Valid Dataset	Test Dataset
0	891	262	126
1	932	259	135
Total	1823	521	261

Figure 1: Methodology Strategy



Tables 1 and 3 show some examples of the tweets that the datasets respective to the Task 1 and 2 contain. Tables 2 and 4 shows the distribution of the classes for each task, after the split into training, valid, and test.

4. Methodology

Addressing the challenges of the competition, we adopted a systematic methodology, comprising several pivotal steps. Given that the competition dataset is in Spanish, our approach primarily leveraged pretrained models tailored for the Spanish language. Additionally, the computations were performed using an NVIDIA RTX 4070 GPU, ensuring efficient processing and model training. Figure 1 illustrates the strategy followed for tasks resolutions.

4.1. Data Splitting By Source

First, the provided dataset was split into two distinct sources: Detest and Stereohoax. This division allowed for handling and analyzing the data in a specific manner according to their origin.

4.2. Split Data By Annotators

Since we opted for the Learning With Disagreement strategy, we subdivided each dataset into subsets according to the annotators from stereotype_a1, stereotype_a2, and stereotype_a3 for the first subtask, and on the annotators implicit_a1, implicit_a2, and implicit_a3 for the second subtask. This segmentation facilitated a structured approach to address annotator discrepancies across both tasks.

4.3. Models

As the text are in Spanish, we decided to use only pre-trained models in Spanish language. The pre-trained models selected, obtained from the Huggingface transformers library, were: [1]

- dccuchile/bert-base-spanish-wwm-uncased [15]. This model (BETO) is a BERT Spanish version
- PlanTL-GOB-ES/roberta-base-bne [16]. This model is based on the RoBERTa base model and has been pre-trained using the largest Spanish corpus known to date

4.4. Baseline

To compare the results obtained with the different models and strategies developed, a starting point was established using selected pretrained models. Since the optimal hyperparameter values cannot be known in advance, some of the most commonly used values were employed: a batch size of 16, a learning rate of 5e-5, a maximum length of 128, and a weight decay of 0.1. The training datasets were used without any additional processing, i.e., as provided by the competition. Tables 5 and 6 show the baseline results on different models for each task.

Table 5
Baseline Results in Subtask 1

		F1 Score
Source	**Model**	**Baseline**
DETESTS	BETO A1	0.7561
	BETO A2	0.7417
	BETO A3	0.7710
	RoBERTa A1	**0.7668**
	RoBERTa A2	**0.7955**
	RoBERTa A3	**0.7919**
STEREOHOAX	BETO A1	0.8496
	BETO A2	0.8331
	BETO A3	0.8412
	RoBERTa A1	**0.8643**
	RoBERTa A2	**0.8389**
	RoBERTa A3	**0.8580**

[1] https://huggingface.co/

Table 6
Baseline Results in Subtask 2

F1 Score		
Source	**Model**	**Baseline**
DETESTS	BETO A1	0.5016
	BETO A2	0.5137
	BETO A3	0.4797
	RoBERTa A1	**0.5394**
	RoBERTa A2	**0.4942**
	RoBERTa A3	**0.5842**
STEREOHOAX	BETO A1	0.6537
	BETO A2	0.7317
	BETO A3	0.7021
	RoBERTa A1	**0.6532**
	RoBERTa A2	**0.7610**
	RoBERTa A3	**0.6343**

4.5. Data Pre-processing

An exhaustive preprocessing of the data was performed to clean and normalize it. The preprocessing stages included:

- Conversion of uppercase to lowercase: To avoid differences caused by capitalization.
- Removal of mentioned users: Elimination of user mentions preceded by '@'.
- Removal of URLs and links: To clean the text of irrelevant external links.
- Removal of hashtags: Only the '#' symbol was removed to preserve relevant words.
- Removal of emoticons: To eliminate graphical elements that do not contribute to textual

Tables 7 shows the result of applying preprocessing.

Table 7
Example of Pre-processing Text

Original	Permiso de residencia para extranjeras víctimas de violencia de género. #abogado #extranjeria #residencia... URL
Pre-processing	permiso de residencia para extranjeras víctimas de violencia de género. abogado extranjeria residencia ... url

Tables 8 and 9 show the results achieved after processing the texts from the comments. As it can be seen, this preproccessing improved the results obtained with the baselines.

4.6. Data Augmentation

Given the umbalanced in class distribution, we opted to address it by employing the back-translation method to augment the dataset. For the primary task, the scarcity of class 1 instances,

Table 8
Pre-processing Results in Subtask 1

		F1 Score	
Source	**Model**	**Baseline**	**Preprocessing**
DETESTS	BETO A1	0.7561	0.7621
	BETO A2	0.7417	0.7547
	BETO A3	0.7710	0.7810
	RoBERTa A1	0.7668	**0.7698**
	RoBERTa A2	0.7955	**0.8037**
	RoBERTa A3	0.7919	**0.7706**
STEREOHOAX	BETO A1	0.8496	0.8591
	BETO A2	0.8331	0.8401
	BETO A3	0.8412	0.8419
	RoBERTa A1	0.8643	**0.8810**
	RoBERTa A2	0.8389	**0.8667**
	RoBERTa A3	0.8580	**0.8647**

Table 9
Pre-processing Results in Subtask 2

		F1 Score	
Source	**Model**	**Baseline**	**Preprocessing**
DETESTS	BETO A1	0.5016	0.5115
	BETO A2	0.5137	0.5235
	BETO A3	0.4797	0.4816
	RoBERTa A1	0.5394	**0.5404**
	RoBERTa A2	0.4942	**0.5062**
	RoBERTa A3	0.5842	**0.6042**
STEREOHOAX	BETO A1	0.6536	0.6736
	BETO A2	0.7317	0.7517
	BETO A3	0.7021	0.7121
	RoBERTa A1	0.6532	**0.6732**
	RoBERTa A2	0.7610	**0.7880**
	RoBERTa A3	0.6343	**0.6745**

the minority class, was a notable challenge across both datasets. To address this, we undertook data augmentation specifically for this class. In the case of the second task, class imbalance was evident in the Detests dataset, where class 0 data was notably sparse. Conversely, the Stereohoax dataset exhibited balance, except for the third annotator's class, which lacked any class 1 instances. To mitigate this issue, we integrated gold label elements containing a 1 in the implicit class, as provided by the competition, into the third annotator's dataset, thereby achieving improved data balance.

This translation process involved initial translation from Spanish to English, then from

English to German, and finally back to English and Spanish, respectively. We utilized the pre-trained model 'Helsinki-NLP/opus-mt-es-en'[17] for the initial translation and 'Helsinki-NLP/opus-mt-en-es' [18] for the reverse translation. Table 10 illustrates the application of this technique to the dataset.

Table 10
Back-Translation Technique

Original	Váyase usted a la m13rd4
First Translation	Go fuck yourself4
Second Translation	Fick dich selbst4
Third Translation	Fuck you4
Back-Translation	Que te jodan4

Tables 11 and 12 the results achieved after applying back-translation technique to the texts from the comments.

Table 11
Back-Translation Results in Subtask 1

		F1 Score		
Source	**Model**	**Baseline**	**Preprocessing**	**Back-translation**
DETESTS	BETO A1	0.7561	0.7621	0.7920
	BETO A2	0.7417	0.7547	0.7660
	BETO A3	0.7710	0.7810	0.7846
	RoBERTa A1	0.7668	**0.7698**	0.7685
	RoBERTa A2	0.7955	**0.8037**	0.8022
	RoBERTa A3	0.7919	**0.7706**	0.7569
STEREOHOAX	BETO A1	0.8496	0.8591	0.7685
	BETO A2	0.8331	0.8401	0.8022
	BETO A3	0.8412	0.8419	0.7569
	RoBERTa A1	0.8643	**0.8810**	0.8462
	RoBERTa A2	0.8389	**0.8667**	0.8437
	RoBERTa A3	0.8580	**0.8647**	0.8297

4.7. Hyperparameter Search

Hyperparameter search is an essential step in fine-tuning models to fit datasets optimally. For this reason, multiple training and testing iterations were performed, using various combinations of hyperparameters such as batch size, learning rate, max size, and weight decay. To minimize training time, datasets were reduced to 80% of their original size before experimentation began. Optuna[19] was the platform used for this process. An exhaustive search algorithm was developed, testing all possible combinations of hyperparameters (grid search). Table 13 shows the search space and the best values found for each of them. It should be noted that the same hyperparameter values were obtained for all trained models.

Table 12
Back-Translation Results in Subtask 2

		F1 Score		
Source	**Model**	**Baseline**	**Preprocessing**	**Back-translation**
DETESTS	BETO A1	0.5016	0.5116	0.5516
	BETO A2	0.5137	0.5237	0.5337
	BETO A3	0.4797	0.4807	0.4997
	RoBERTa A1	0.5394	0.5404	**0.5694**
	RoBERTa A2	0.4942	0.5062	**0.5142**
	RoBERTa A3	0.5842	0.6042	**0.6242**
STEREOHOAX	BETO A1	0.6536	0.6736	0.6936
	BETO A2	0.7317	0.7518	0.7718
	BETO A3	0.7021	0.7121	0.7321
	RoBERTa A1	0.6532	0.6734	**0.6932**
	RoBERTa A2	0.7610	0.7881	**0.8000**
	RoBERTa A3	0.6343	0.6745	**0.6843**

Table 13
Hyperparameter Space and Best Hyperparameter Values

Hyperparameter	**Values**
Batch size	[16, **32**]
Learning rate	[**3e-05**, 5e-05]
Max length	[64, **128**]
Weight decay	[0.01, **0.1**]

The hyperparameter search was used on the best-performing models, namely, preprocessed BETO and RoBERTa, as they yielded the best results. Table 14 displays the outcomes of the hyperparameter search. The Performance column presents the results of training with the best hyperparameter found on the preprocessed datasets.

4.8. Ensemble Approach

Finally, a classification model was developed by combining the three models trained for each annotator using a hard voting approach. This ensemble method improved overall performance by reducing individual biases from each annotator and providing more robust predictions. Table 15 and 16 present the chosen model for the ensemble; for both datasets, the RoBERTa models were selected as they produced the best results. Additionally, a comparison is made with the results obtained from training the model with the gold label provided by the competition. It is evident that the learning with disagreement approach yielded superior results. As can be observed, applying LeWiDi has clearly been a successful approach.

Table 14
Hyperparameter Search Results in Subtask 1

		F1 Score			
Source	**Model**	**Baseline**	**Preprocessing**	**Back-translation**	**Performance**
DETESTS	BETO A1	0.7561	0.7621	0.7920	0.7970
	BETO A2	0.7418	0.7548	0.7660	0.7670
	BETO A3	0.7711	0.7810	0.7847	0.7878
	RoBERTa A1	0.7668	0.7698	0.7685	**0.7729**
	RoBERTa A2	0.7955	0.8038	0.8023	**0.8080**
	RoBERTa A3	0.7919	0.7706	0.7569	**0.7846**
STEREOHOAX	BETO A1	0.8496	0.8591	0.8454	0.8687
	BETO A2	0.8331	0.8402	0.8269	0.8553
	BETO A3	0.8412	0.8419	0.8073	0.8137
	RoBERTa A1	0.8643	0.8811	0.8463	**0.8924**
	RoBERTa A2	0.8389	0.8668	0.8437	**0.8696**
	RoBERTa A3	0.8580	0.8647	0.8297	**0.8649**

Table 15
Comparison of Gold Label vs Ensemble in Subtask 1

		F1 Score	
Source	**Model**	**Gold Label**	**Ensemble**
DETESTS	RoBERTa	0.7899	**0.7949**
STEREOHOAX	RoBERTa	0.8896	**0.8973**

Table 16
Comparison of Gold Label vs Ensemble in Subtask 2

		F1 Score	
Source	**Model**	**Gold Label**	**Ensemble**
DETESTS	RoBERTa	0.5633	**0.5800**
STEREOHOAX	RoBERTa	0.7096	**0.7246**

5. Results

To measure our results, the organizers provided an unlabeled test dataset, which we processed with our models to generate predictions for each instance. These predictions were then used to calculate our score for the leaderboard. In the first task, we achieved the second position (I2C-Huelva_1) using an ensemble consisting of 3 models, one for each annotator, based on RoBERTa. A second run (I2C-Huelva_2), consisting of an ensemble of three models, one for each annotator, based on BETO, was submitted and achieved third place with an F1 Score of 0.701. Table 17 shows the ranking summary.

For Task 2, the submitted run consisted of an ensemble of three models, one for each annotator,

Table 17
Ranking Results for the Subtask 1

Ranking	User	F1-Score
1	Brigada Lenguaje_1	0.724
2	**I2C-Huelva_1**	**0.712**
3	**I2C-Huelva_2**	**0.701**
4	EUA_2	0.691
-	-	-
20	BASELINE_fast_text_svc	0.297

based on RoBERTa. This run achieved 7th place with an ICM of -0.328. Table 18 shows the ranking summary:

Table 18
Ranking Results for the Subtask 2

Ranking	User	ICM	ICM Norm
1	BASELINE_beto	0.126	0.546
2	EUA_2	0.065	0.524
3	EUA_3	0.061	0.522
4	EUA_1	0.045	0.516
5	Brigada Lenguaje_1	-0.240	0.413
6	BASELINE_tfidf_svc	-0.275	0.400
7	**I2C-Huelva_1**	**-0.328**	**0.381**
-	-	-	-
14	UC3M-SAS_2	-2.103	0.000

6. Error Analysis

For error analysis, our test dataset built during the development phase has been used. The confusion matrices of the classifiers for both tasks can be found in Figures 2, 3, 4 and 5. Figure 2 shows the performance of the classifier in predicting classes Gold 0 and Gold 1 for the Detests task. While the classifier accurately predicts most instances of the Gold 0 class (392 true negatives), it is less reliable in predicting the Gold 1 class, with only 89 true positives and 47 false negatives. This outcome suggests that the classifier tends to predict the Gold 0 class more accurately, likely due to an imbalance in the training dataset where the Gold 1 class is underrepresented.

Figure 3 illustrates the classifier's performance in the Stereohoax task. In this instance, the classifier demonstrates a better balance between predicting both classes. It correctly predicts 287 instances of the Gold 0 class and 105 instances of the Gold 1 class. However, some errors persist, with 16 false positives and 20 false negatives. This more balanced performance may be attributed to a more even distribution of classes in the training dataset.

Figure 2: Confusion Matrix Task 1 Detests

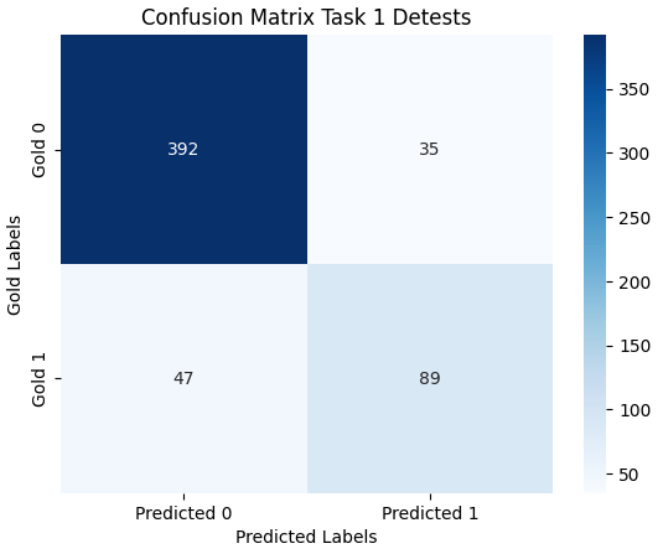
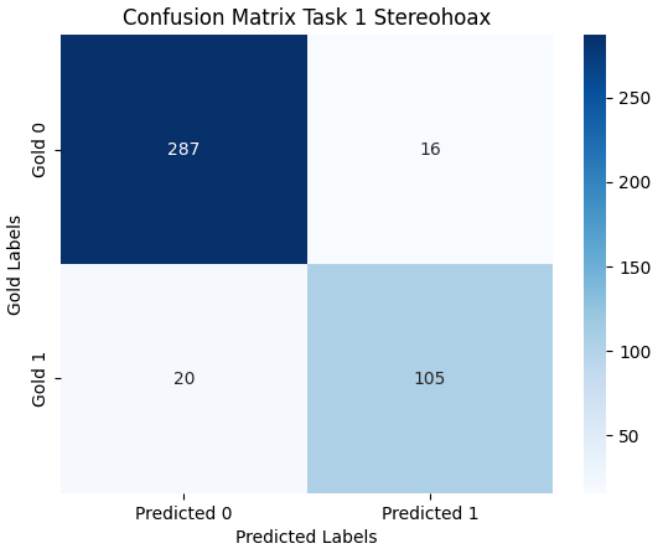


Figure 3: Confusion Matrix Task 1 Stereohoax



To analyze Task 2, only the cases classified as positive in Task 1 have been used. Therefore, the analysis was conducted as a binary task to distinguish between the explicit and implicit classes.

Figure 4 displays the confusion matrix for Task 2 of Detests dataset. Similar to the previous results, the classifier shows a strong performance for the Gold 0 class, with 87 true negatives and 15 false positives. However, the performance for the Gold 1 class remains weak, with 15 true positives and 8 false negatives. This indicates that the class imbalance issue observed in Task

Figure 4: Confusion Matrix Task 2 Detests

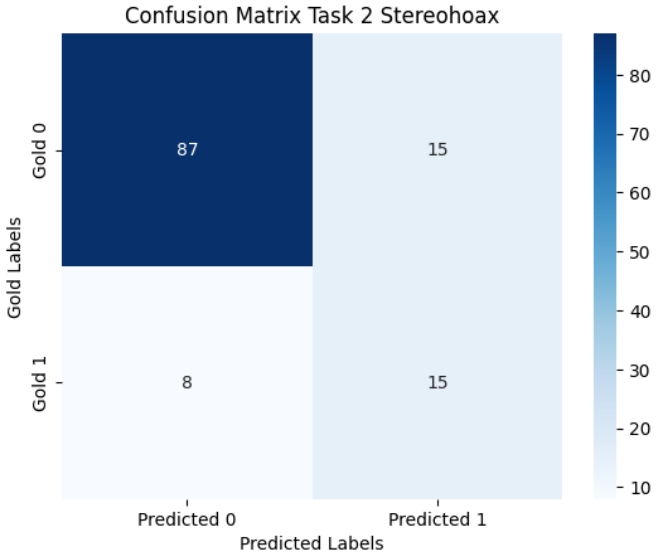
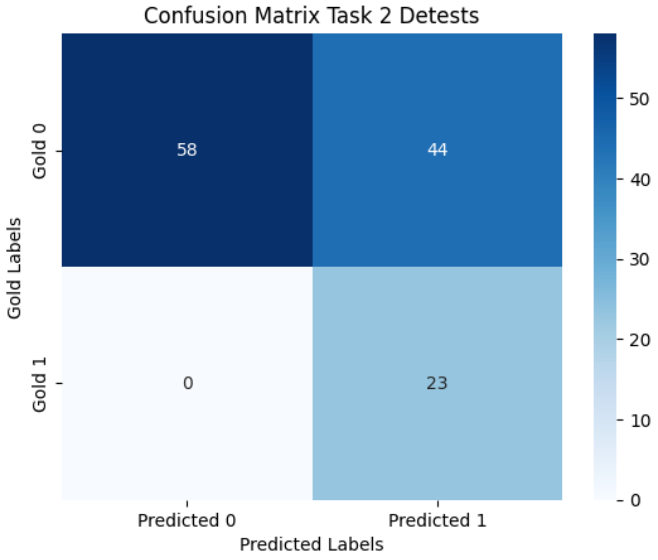


Figure 5: Confusion Matrix Task 2 Stereohoax



1 continues to affect the classifier's performance in this task. Figure 5 presents the confusion matrix for Task 2 of Stereohoax. Here, the classifier exhibits high performance in predicting the Gold 0 class with 87 true negatives but struggles with predicting the Gold 1 class, yielding only 15 true positives and 8 false negatives. This pattern suggests that the classifier may be influenced by a class imbalance similar to that observed in Task 1 of Stereohoax.

7. Conclusion

In this competition, we introduced the DETESTS-Dis task as part of IberLEF 2024, aimed at detecting and classifying explicit and implicit stereotypes in texts from social media and comments on news articles, incorporating learning with disagreement techniques. By leveraging the annotations provided by multiple annotators and employing learning with disagreement, we successfully mitigated individual biases and uncertainties inherent in the data labeling process. Furthermore, through the ensemble technique, we combined the predictions from each annotator's model and obtained a final prediction using a majority voting approach, which proved to be highly effective in enhancing the overall predictive performance.

Additionally, in Task 1, we achieved an F1-Score of 0.712, reflecting a robust performance in stereotype identification. In Task 2, we obtained an ICM of -0.328 and a normalized ICM of 0.381, indicating a detailed analysis of the presence and manifestation of stereotypes, both explicit and implicit. These results highlight our dedication to improving stereotype detection techniques and tackling the complex challenges posed by social media and news commentary.

Acknowledgments

This paper is part of the I+D+i Project titled "Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF/EU".

References

[1] F. H. Allport, The structuring of events: outline of a general theory with applications to psychology., Psychological Review 61 (1954) 281.

[2] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, A. Ariza-Casabona, M. Farrús, P. Rosso, M. Taulé, Overview of DETESTS-Dis at IberLEF 2024: DETEction and classification of racial STereotypes in Spanish - Learn with Disagreement, Procesamiento del Lenguaje Natural 69 (2024).

[3] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, 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, 2024.

[4] E. Fersini, D. Nozza, P. Rosso, et al., Overview of the evalita 2018 task on automatic misogyny identification (ami), in: CEUR Workshop Proceedings, volume 2263, CEUR-WS, 2018, pp. 1–9.

[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[6] P. Chiril, F. Benamara, V. Moriceau, "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, 2021, pp. 2833–2844.
[7] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, B. Y. Zhao, Detecting gender stereotypes: Lexicon vs. supervised learning methods, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–11.
[8] P. Vossen, A. Fokkens, I. Maks, C. van Son, Towards an open dutch framenet lexicon and corpus, in: Proceedings of the LREC 2018 Workshop International FrameNet Workshop, 2018, pp. 75–80.
[9] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, arXiv preprint arXiv:1911.03891 (2019).
[10] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).
[11] J. Sánchez-Junquera, P. Rosso, M. Montes, B. Chulvi, et al., Masking and bert-based models for stereotype identication, Procesamiento del Lenguaje Natural 67 (2021) 83–94.
[12] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish, Procesamiento del lenguaje natural 69 (2022) 217–228.
[13] E. Leonardelli, C. Casula, Dh-fbk at semeval-2023 task 10: Multi-task learning with classifier ensemble agreement for sexism detection, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1894–1905.
[14] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.
[15] J. Cañete, G. Chaperon, R. Fuentes, J. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: Proceedings of the Workshop on Practical Machine Learning for Developing Countries (PML4DC) at ICLR 2020, 2020.
[16] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022).
[17] J. Tiedemann, S. Thottingal, Opus-mt: Building open translation services for the world, https://huggingface.co/Helsinki-NLP/opus-mt-es-en, 2020. Accessed: 2024-07-08.
[18] J. Tiedemann, S. Thottingal, Opus-mt: Building open translation services for the world, https://huggingface.co/Helsinki-NLP/opus-mt-en-es, 2020. Accessed: 2024-07-08.
[19] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.