FairTransNLP: Fairness and Transparency for Equitable NLP Applications in Social Media

Paolo Rosso[1,2], Mariona Taulé[3], Laura Plaza[4] and Jorge Carrillo-de-Albornoz[4]

[1] *Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*
[2] *ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence, Spain*
[3] *Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona, Gran Via 585, 08029, Barcelona, Spain*
[4] *Universidad Nacional de Educación a Distancia, Juan del Rosal, 16, 28040 Madrid, Spain*

Abstract

Artificial Intelligence (AI) applications often perpetuate and accentuate unfair biases that can originate from multiple sources, such as data sampling, labelling and training data. Biased outputs can negatively affect certain social groups of users and even lead to discrimination. The ability of AI systems to provide transparent and understandable explanations for their decisions is crucial both for developers, to better understand the systems' behaviour, and for users, to gain trust in AI systems. In this coordinated project (UPV, UB, UNED), we have addressed problems such as the detection and classification of racial stereotypes and sexism in social networks, considering the multiple perspectives of annotators in data that have "conflicting" labels. This was made possible by employing the Learning with Disagreements paradigm with the aim to foster the development of more equitable AI models, i.e. fair and inclusive towards multiple viewpoints, rather than only representing the majority view.

Keywords

Fairness, transparency, equitable NLP, learning with disagreements, racial stereotypes, sexism

1. Motivation and Related Work

Biased systems are those that systematically and unfairly discriminate against individuals or social groups. If a biased system becomes widely adopted, the social biases it perpetuates may have serious consequences. Another major issue is biases in the design and annotation of datasets. For example, in [1], the authors point out how annotated data may carry racial biases, and how models can learn such biases. Topic bias is another factor to consider when developing datasets. Recent studies showed how the volatile nature of topics, especially on social media, can hinder the predictive capability of models trained on data collected with keyword sets [2] or in restricted time spans. In [3] the authors analyse the topic bias in datasets used in shared tasks on the detection of toxic/abusive language, hate speech and offensive language, misogyny and sexism.

In the framework of the FairTransNLP project, we have analysed bias in several domains: (i) media bias, (ii) hate speech and (iii) multimodal sexism. The problem of media bias was reviewed in [5]. In that study, the authors concluded that the current methods for automatic media bias detection are still in their infancy and there is still a lot of potential for improvement in terms of accuracy and robustness. In another work [6], the same authors employed LIME and SHAP explainability techniques to determine to what extent lexical-based AI models could identify bias. In [7], bias was studied in pre-trained models for hate speech detection, showing that they are biased towards hateful keywords: fine-tuning these models with hateful texts that do not contain the hateful keywords making possible to reduce the bias. Finally, in [8], a bias estimation technique was proposed to identify specific elements that compose a meme that could lead to unfair models, together with a bias mitigation strategy based on Bayesian Optimization.

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

prosso@dsic.upv.es (P. Rosso); mtaule@ub.edu (M. Taulé); mfarrus@ub.edu (M. Farrús); plaza@lsi.uned.es (L. Plaza); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz)

0000-0002-8922-1242 (P. Rosso); 0000-0003-0089-940X (M. Taulé); 0000-0002-7160-9513 (M. Farrús); 0000-0001-5144-8014 (L. Plaza); 0000-0002-1449-1547 (J. Carrillo-de-Albornoz)

In the next sections, after describing Learning with Disagreements, we will comment on two problems we addressed employing this new paradigm: the detection and classification of racial stereotypes and sexism in social networks. The aim was to foster the development of more fair and inclusive AI models in the framework of two shared tasks we organized at **IberLEF** (racial stereotypes) and **CLEF** (sexism) forums.

2. Learning with Disagreements

The classic approach for dealing with disagreement among annotators assumes the existence of a single, objective label, known as the gold label, which can be extracted through a majority voting scheme. Although this methodology is simple and effective, it is also true that it ignores the opinion of the minority over the majority, hence neglecting other viewpoints. In fact, disagreement can be considered as a signal, not as noise [9], as it provides useful information for learning [10].

For instance, consider the tweet "*Women should stop trying to understand football and focus on what they're good at, like fashion.*" This example illustrates the use of soft labels, in which annotator disagreement is preserved rather than collapsed into a single category. Out of 6 annotators, 4 labeled the tweet as sexist and 2 as non-sexist, resulting in a soft label distribution of *sexist: 0.67, non-sexist: 0.33.*

Learning With Disagreements (LeWiDi) usually applies a soft loss approach. That means that, although it also aggregates the annotations of each annotator, it aggregates them into a probability distribution. The main goal behind this approach is to optimize a model distribution to resemble the original one produced by disagreement among annotators [11]. In contrast, the perspectivist approach disregards aggregation and proposes working directly with individual annotations [12]. Modelling strategies for LeWiDi include using soft labels that reflect the distribution of annotator responses, modeling individual annotators to capture biases and reliability, and applying multi-task or probabilistic approaches to jointly infer true labels and annotator behavior.

One of the key conclusions drawn from the LeWiDi-inspired tasks EXIST and DETEST, is the clear performance improvement achieved when training systems with soft labels instead of hard labels. This finding, which aligns with previous literature, highlights how converting multiple annotator judgments into a single ground truth label leads to a loss of valuable information—both in terms of discarded instances (to resolve ties or disagreements) and in the reduction of nuanced perspectives within each instance. Moreover, commonly used disagreement resolution methods (e.g., majority vote, adding an extra annotator) are often applied without considering the specifics of the task, use case, or annotator profiles, which unintentionally introduce biases into the final dataset.

3. DETESTS: DETEction and classification of racial STereotypes in Spanish

The DETESTS-Dis corpus was created with the aim of analyzing and detecting stereotypes related to immigration on social media. We adopted the LeWiDi paradigm given the inherent subjectivity of this task. The identification of stereotypes on social media is crucial because their presence reinforces toxic and hate speech against vulnerable social groups such as immigrants. Furthermore, these types of messages are rapidly disseminated on social media such as Twitter (currently X). Stereotype detection is a complex task both because of its subjective nature (the same message can be interpreted differently depending on the culture, beliefs, age or gender of the reader), and because of the way in which stereotypes can be expressed, i.e. explicitly or implicitly. The different human perspectives and stereotypes implicit in the messages (i.e., when a certain inference process is required to understand them) are possibly the main difficulties for detection systems. Based on these assumptions, we created the **DETESTS-Dis** dataset, which consists of two corpora from different social media sources containing two types of texts:

The **StereoCom** corpus consists of 6,762 sentences extracted from 3,054 comments posted in response to articles manually selected from 12 Spanish online newspapers, including El País, La Vanguardia and ABC and discussion forums like Menéame and ForoCoches, related to immigration. This selection was carried out considering news articles published between August 2017 and November 2021 containing controversial content, potential toxicity and a minimum of 50 published comments per article, following the methodology applied to the NewsCom-TOX corpus [13]. We used a keyword-based approach to search for articles mainly related to xenophobia likely to include ethnic stereotypes related to immigrants.

The **SteroHoax-ES** corpus consists of 5,349 Twitter messages retrieved in 2021 from 449 conversational heads (i.e., the tweet starting the conversation) responding to 72 racist hoaxes related to immigration in Spain, manually extracted from the fact-checking websites Maldita.es and Newtral, which verify or refute claims made on social networks. The tweets were retrieved using keywords and the contents of the hoaxes with Twitter API v2 for Academia. We also collected the conversational threads. This corpus corresponds to the Spanish subset of the StereoHoax multilingual dataset [14]. Both corpora were annotated with the following categories:

- **Stereotype**: a binary category indicating the presence or absence of stereotypes.
- **Stereotype classification**: a multilabel category in which immigrants are presented as: (*1)* '*victims of xenophobia*', *(2)* '*suffering victims*', *(3)* '*economic resources*', *(4)* a problem of '*migration control*', *(5)* people with '*cultural and religious differences*', *(6)* people who receive the '*benefits*' of our social policy, *(7)* a problem for '*public health*', *(8)* a threat to '*security*', *(9)* '*dehumanization*' and *(10)* '*other*' types of stereotypes. This classification is based on the proposal of [15].
- **Implicitness**: a binary category indicating whether the stereotype is expressed explicitly or implicitly.
- **Type of Implicitness**: a multilabel category including different linguistic strategies used to convey implicit stereotypes: *(1)* '*world knowledge'*; *(2)* '*figures of speech'* (such as metaphor, rhetorical questions, euphemisms and reported speech); *(3)* '*humor/jokes'*; *(4)* '*irony/sarcasm*'; *(5)* '*extrapolation*'; *(6)* '*imperative/exhortative calls*' for action related to immigrants; *(7)* '*entailment/ evaluation*'; and the *(8)* '*others*' label for types of implicitness not considered in the previous categories. We also included the *(9)* 'context' label to indicate that it is necessary to consider previous messages (sentences or tweets) to understand the implicit stereotype.

All these corpora were annotated by three linguists of different ages and genders (one expert linguist and two trained students). The DETESTS-Dis dataset was released both in its aggregated form (applying the majority vote, hard labels) and its disaggregated form (soft labels), in the DETESTS-Dis task [16], which took place as part of **IberLEF 2024**. The conversational threads of comments and tweets were also provided to the participants in the task, but not the information included in the stereotype classification and type of implicitness labels. The DETESTS-Dis task (DETEction and classification of racial Stereotypes in Spanish - Learning with Disagreement) was designed hierarchically by chaining two binary-classification subtasks:

1. The **stereotype detection subtask** aimed to determine whether a tweet or sentence contains any stereotype, considering the full distribution of labels provided by the annotators: (*1)* ***Stereotype****: […] Illegal immigrants have more rights than Spaniards and we are FED UP! (2)* ***Not Stereotype****: In fact, all Muslim-majority countries have Sharia as a source of law to a greater or lesser degree, they are all theocratic.*
2. The **implicitness identification subtask** introduces a hierarchical binary classification problem to identify whether the stereotypes in the text are explicit or implicit: *(1)* ***Implicit****: Quality immigrants. (2)* ***Explicit****: What is dangerous is the brainless immigrants, who misinterpret Islam and who kill innocent people.*

The first subtask was evaluated using the binary F1 metric for the models that output hard labels, and the cross-entropy metric was applied between the system soft label values and the soft labels generated from the average votes of the annotators. The ICM metric [17], an information theoretic-based metric that considers both the hierarchical structure and the class specificity, was used to evaluate the second subtask. The ICM metric was the official metric for the ranking of both hard labels (ICM) and soft labels (ICM-Soft). 15 teams signed up to participate, six of whom sent runs, while three sent a working paper. The models used by participants were BETO, RoBERTa, XML-RoBerta, Twitter-RoBERTa and Twitter-XML-RoBERTa while one of them used GPT-3.5-Turbo. Various teams employed data augmentation techniques, mainly back-translation. Only a few teams used contextual information. The best system in subtask1 with hard labels achieved a 0.724 F1, while soft labels achieved a Cross Entropy score of 0.841. No team beat the BETO baseline with hard labels and only one team achieved a better result with soft labels (ICM-soft normalized of 0.403).

4. EXIST: sEXism Identification in Social neTworks

EXIST is a series of scientific events and shared tasks aimed at capturing sexism in a broad sense, from explicit misogyny to subtle, implicit sexist behaviours. The last two editions were held as labs at **CLEF 2023** and **CLEF 2024**, while the first two took place at **IberLEF**.

The EXIST shared task focuses on identifying and classifying sexism in social networks. While the 2023 edition focused on tweets, the 2024 edition expanded the task scope by incorporating memes, recognizing their increasing use in spreading harmful messages disguised as humour. Both editions, however, included the three main following tasks:

- **Sexism Detection:** Binary classification to determine whether a content is sexist or not.
- **Source Intention Classification:** Ternary classification distinguishing between:
  - ***Direct messages:*** *Women shouldn't code... perhaps be an influencer instead...it's their natural strength.*
  - ***Reported messages:*** *Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.*
  - ***Judgmental messages:*** *As usual, the woman was the one quitting her job for the family's welfare...*
- **Sexism Categorization:** Multi-label classification assigning sexist content to one or more among five categories:
  - ***Ideological and inequality:*** *#Feminism is a war on men, but it's also a war on women.*
  - ***Role stereotyping and dominance:*** *I feel like everytime I flirt with a girl they start to imagine all the ways they can utilize me.*
  - ***Objectification:*** *No offense but I've never seen an attractive african american hooker. Not a single one.*
  - ***Sexual violence:*** *Fuck that cunt, I would with my fist.*
  - ***Misogyny and non-sexual violence:*** *Domestic abuse is never okay.... Unless your wife is a bitch.*

Both editions adopted the **LeWiDi paradigm** and conducted both **hard and soft evaluations** [18]. The hard evaluation assigned discrete labels to instances and used ICM [17] and F1 as metrics, while the soft evaluation assessed the model's ability to capture label disagreement by comparing probability distributions using ICM-Soft and Cross Entropy.

The **2023 edition of EXIST** focused on textual data from microblogs. The dataset consisted of over 5,000 tweets in English and Spanish**,** collected using more than 200 potentially sexist phrases sourced from academic studies, tweets from journalists and activists reporting sexist incidents,

expressions from Everyday Sexism[1], and feminist dictionaries. Rather than relying on majority voting, each tweet was labeled by multiple annotators from diverse socio-demographic backgrounds, ensuring a broad representation at the gender, age, country and education levels. Both in EXIST 2023 and EXIST 2024, all social media data was anonymized to prevent the identification of individuals, and data collection adheres strictly to the terms of service of the respective platforms. During annotation, participants were explicitly warned that the content may include sensitive material and are free to withdraw at any time. Furthermore, we relied on platforms such as Prolific, which require annotators to comply with ethical guidelines.

A total of **28 teams from 29 countries submitted 232 runs**. For a comprehensive description, please refer to the overview of the task [19]. Approximately 90% of the systems relied on LLMs. Only a few teams explored traditional machine learning methods. Several participants applied data augmentation techniques, such as tweet translation, external datasets, and instance duplication. Twitter-specific models and transfer learning from related domains, such as hate speech, toxicity, and sentiment analysis, were used. The best-performing system submitting soft labels achieved an ICM-soft normalized score of 0.6421, 0.8072, and 0.7879, respectively, for the three tasks. For systems submitting hard labels, the best F1 scores were 0.8109, 0.5715, and 0.6296, respectively.

The **2024 edition of EXIST** [19] expanded the task to multimodal content by introducing memes. The dataset includes both the EXIST 2023 tweet collection and a newly curated set of memes. To retrieve relevant memes, 250 sexist-related terms were used as search queries on Google Images, obtaining the top 100 images per term. A manual cleaning process was applied to remove irrelevant content such as ads, duplicates, textless images, and text-only images. The final dataset contains 2,000 memes per language for training and 500 memes per language for testing. The following figures show examples of sexist memes.

Figure 1: Examples of sexism memes

A total of **57 teams submitted 412 runs**, marking a significant increase in participation. For a comprehensive description, please refer to the overview of the task [20]. Most teams relied on monolingual and multilingual LLMs (BERT, DistilBERT, MarIA, MDEBERTA, RoBERTa, DeBERTa, LLaMA, and GPT-4). For meme analysis, vision models such as CLIP, BEIT, and VIT were employed. Some teams integrated linguistic features, while others used data augmentation and prompt engineering. A small number of participants explored deep learning architectures or traditional ML methods. As in EXIST 2023, Twitter-specific models were used.

For systems submitting soft labels, the best ICM-soft normalized scores were 0.6755 (tweets) - 0.4530 (memes), 0.4795 (tweets) - 0.3676 (memes), and 0.4379 (tweets) - 0.2462 (memes) for the three tasks, respectively. For systems submitting hard labels, the best F1 scores were 0.7944 (tweets) - 0.7642 (memes), 0.5677 (tweets) - 0.3873 (memes), and 0.6004 (tweets) - 0.4319 (memes). As can be seen, identifying and characterizing sexism in memes seems more difficult than in text.

5. Conclusions and Future Work

This work shows that incorporating annotation disagreement improves NLP models on subjective tasks like detecting sexism and stereotypes. Results from DETESTS and EXIST confirm that soft labeling using annotator distributions yields better performance and richer representations than majority voting, which can mask interpretive differences and reinforce bias.

[1] https://everydaysexism.com/

As future work, in EXIST 2025 [20], we will analyze sexist content in TikTok videos, addressing the influence of short-form platforms on stereotype diffusion. This will introduce new challenges in annotation, multimodal modeling, and fairness-aware evaluation.

Acknowledgements

This work has been funded by MCIN/AEI/10.13039/501100011033 and ERDF/EU (PID2021-124361OB-C31-C32-C33).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The Risk of Racial Bias in Hate Speech Detection, in: Proceedings of the 57th Annual Meeting of the ACL, 2019, pp. 1668–1678.

[2] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: The Problem of Biased Datasets, in: Proceedings of the 2019 Conf. of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1), 2019, pp. 602–608.

[3] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media, Applied Sciences 10(12) (2020) 4180.

[4] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review, Lang. Resour. Eval. 55(2) (2021) 477–523.

[5] F-J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, L. Plaza, A Systematic Review on Media Bias Detection: What is Media Bias, How it is Expressed, and How to Detect it, Expert Syst. Appl. 237 (2024) 121641.

[6] F-J. Rodrigo-Ginés, J. Carrillo-de-Albornoz, L. Plaza, Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection. Procesamiento del Lenguaje Natural 71 (2023) 179–190.

[7] G. De La Peña, P. Rosso, Systematic Keyword and Bias Analyses in Hate Speech Detection, Inf. Process. Manag. 60(5) (2023) 103433.

[8] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing Misogynous Memes: Biased Models and Tricky Archetypes, Inf. Process. Manag. 60(5) (2023) 103474.

[9] L. Aroyo, C. Welty, Truth is a Lie: Crowd Truth and the Seven Myths of Human Annotation, AI Magazine 36(1) (2015) 15–24.

[10] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from Disagreement: A Survey. J. Artif. Intell. Res. 72 (2021) 1385–1470.

[11] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A Case for Soft Loss Functions, in: Proceedings of the AAAI Conf. on Human Computation and Crowdsourcing (Vol. 8), 2020, pp. 173–177.

[12] F. Cabitza, A. Campagner, V. Basile, Toward a Perspectivist Turn in Ground Truthing for Predictive Computing, in: Proceedings of the 37th AAAI Conf. on Artificial Intelligence, AAAI'23, 2023.

[13] M. Taulé, M. Nofre, V. Bargiela, X. Bonet, NewsCom-TOX: A Corpus of Comments on News Articles Annotated for Toxicity in Spanish, Language Resources and Evaluation 58 (2024) 1115–1155.

[14] W.S. Schmeisser-Nieto, A.T. Cignarella, T. Bourgeade, S. Frenda, A. Ariza-Casabona, M. Laurent, P.G. Cicirelli, A. Marra, G. Corbelli, F. Benamara, C. Bosco, V. Moriceau, M. Paciello,

M. Taulé, F. D'Errico, Stereohoax: A Multilingual Corpus of Racial Hoaxes and Social Media Reactions Annotated for Stereotypes, Lang. Resour. and Eval. 59 (2025) 2031–2069.

[15] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S.P. Ponzetto, How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants, Applied Sciences 11(8) (2021) 3610.

[16] 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 - Learning with Disagreement, Procesamiento del Lenguaje Natural 73 (2024) 323–333.

[17] E. Amigo, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), 2022, pp. 5809–5819.

[18] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), Working Notes of the Conference and Labs of the Evaluation Forum (2023).

[19] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), Working Notes of the Conference and Labs of the Evaluation Forum (2024).

[20] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: Proceedings of the European Conference on Information Retrival, ECIR, 2025, pp. 442-449.