Results in Engineering 30 (2026) 110202

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Capturing subjectivity: A weighted ensemble approach to preserve annotator diversity

Laura Vázquez Ramos *, Jacinto Mata Vázquez , Victoria Pachón Álvarez

I2C Research Group, CITES Research Center, Information Technology Department, Universidad de Huelva, Spain

ARTICLE INFO

Keywords:
Learning with disagreement (LeWiDi)
Weighted ensemble
Natural language processing
Annotator diversity
Sexism
Sentiment analysis

ABSTRACT

Subjective linguistic tasks, such as sexism detection or sentiment analysis, often involve substantial disagreement among human annotators, reflecting genuine interpretive diversity rather than annotation noise. Traditional aggregation methods, most commonly majority voting, enforce a single reference label and an artificial consensus. This is problematic because it discards information about how different groups of people interpret the same content, thereby obscuring nuances that are crucial for understanding the phenomenon under study. This paper introduces a perspectivist framework that explicitly models annotator diversity by training independent classifiers based on demographic variables and subsequently combining them through a weighted ensembling strategy. Each perspective is assigned a relative importance according to its individual performance (F_1-score), and the decision threshold is optimised to maximise the overall F_1-score of the ensemble. Experiments conducted on three datasets-EXIST Texts 2024, EXIST Memes 2024, and a re-annotated version of SST-2-show consistent improvements across all tasks. The weighted ensemble achieves an F_1-score of 0.84 on EXIST Texts, improves performance from 0.84 to 0.91 on EXIST Memes, and attains an F_1-score of 0.95 on the re-annotated SST-2 dataset. These results demonstrate that weighted perspectivist ensembling achieves a better balance between precision and recall than both individual models and standard baselines, while preserving human interpretive diversity. They highlight the potential of perspectivist modelling as a pathway towards fairer and more robust NLP systems that are better aligned with human variability.

1. Introduction

In natural language processing (NLP) tasks that involve a high degree of subjectivity, such as hate speech, irony, sexism or emotion detection, annotator disagreement is not only common but also to be expected. These differences do not necessarily reflect errors or inconsistencies; instead, they capture the diversity of interpretations [1] that a single text can elicit in different individuals, shaped by sociocultural, ideological or experiential factors [2].

Conventionally, the standard practice in dataset construction has been to address these disagreements through aggregation strategies, most commonly majority voting. This approach aims to enforce a single "truth" or artificial consensus, thereby collapsing the task into a single label [3].

However, this process discards valuable information about the inherent variability of human judgement, which potentially limits models when handling ambiguous or borderline cases. In recent years, several lines of research have argued for preserving and exploiting disagreement rather than discarding it. In particular, the Learning with Disagreement (LeWiDi) paradigm proposes that annotator differences can be informative and useful for improving the robustness and fairness of models [4]. Under this approach, disagreement is treated as knowledge about the multiple perspectives from which a linguistic phenomenon can be interpreted. Various methodologies have been developed to incorporate these divergences into the training process [5]. Existing disagreement-aware methods often focus on modelling multiple annotations directly (e.g., via soft labels or uncertainty-aware objectives) or on learning annotator-specific behaviours. However, they typically do not explicitly operationalise demographic perspectives as independent modelling units, nor do they systematically study how to combine such perspectives through calibrated ensemble strategies. In contrast, our approach defines perspectives using annotator demographic metadata, trains separate models for each subgroup, and integrates them via a performance-based weighted ensemble calibrated on held-out data, allowing complementary signals across perspectives to be combined while attenuating subgroup-specific biases.

Building on this line of work, a methodology is proposed to explore the potential of perspectivist models. These are models trained on the

* Corresponding author.
E-mail address: laura.vazquez@dti.uhu.es (L. Vázquez Ramos).

https://doi.org/10.1016/j.rineng.2026.110202
Received 22 December 2025; Received in revised form 9 March 2026; Accepted 20 March 2026
Available online 27 March 2026

annotations of different groups or individuals, each representing a distinct way of interpreting the same content. They are then combined using a weighted mechanism that captures perspective-specific reliability and bias of each perspective. In doing so, the approach preserves the richness of disagreement while at the same time optimising overall performance by integrating multiple points of view [6,7].

A general framework is presented for applying perspectivist learning to classification tasks involving subjective content. This supports developing models that do not ignore annotation diversity but instead incorporate it as part of the modelling process. Specifically, three fundamental goals are pursued:

- Define a systematic methodology for training independent models based on different annotation perspectives, whether individual or group.
- Implement a weighted combination mechanism that assigns specific weights to each model based on its reliability or suitability for the task, thus optimizing the joint representation of perspectives.
- Evaluate the methodology on three real datasets characterized by disagreement between annotators or by an explicit perspectivist design.

Additionally, this work contributes to the ongoing debate on how subjectivity should be represented in NLP by offering a methodological alternative that preserves interpretative plurality without compromising model accuracy or generalisation.

The study is organised around research questions (RQs) designed to examine in depth the validity and potential of the proposed perspectivist approach. These questions guide the experimental design and the interpretation of results, and focus on assessing the influence of disagreement, the usefulness of the weighted ensemble, and the generalisability of the methodology.

- **RQ1:** Can the use of models trained on different perspectives improve classification in tasks with high levels of disagreement?
- **RQ2:** What impact do the number and diversity of annotators have on the effectiveness of a perspectivist approach?
- **RQ3:** Does the use of perspective-optimized weights improve on simple voting or uniform ensembles?
- **RQ4:** Can the methodology based on perspectivist models and weighted ensembling be successfully applied to other datasets that exhibit disagreement between annotators?

These questions examine whether human diversity in annotation can be turned into a computational advantage, and whether it is possible to design NLP models that not only tolerate subjectivity, but actively learn from it.

2. Related works

Recent studies have addressed annotator disagreement, viewing subjectivity as a valuable source of information rather than noise.

In [8], the authors analyse how NLP systems can reproduce inequalities if the demographic variation present in annotation data is not taken into account. Their work shows that differences in interpretation are not noise, but rather a reflection of unequal experiences and distinct sociocultural frameworks that must be considered when designing fair and representative models. From a methodological view, [5] propose a formal framework for modelling disagreement using probability distributions instead of single labels, and empirically show that preserving multiple perspectives improves both the robustness and the fairness of the models. In a similar direction, [7] employs the *Learning with Disagreement* paradigm for sexism detection in memes, training perspectivist Transformer-based models and combining them through weighted strategies. It confirms that integrating disagreement yields better results than majority voting.

This work [9] provides an extensive analysis of how bias can emerge and propagate in NLP systems. The authors identify five fundamental sources (data, annotators, tasks, models and metrics) and show how each of these can introduce distortions that affect both system performance and fairness. In particular, the article highlights that annotators' decisions and characteristics constitute a critical locus of bias, as their cultural, social and demographic backgrounds directly shape the subjective judgements on which datasets are built.

A particularly relevant study is [10], which systematically analyses how annotator disagreement is not an accidental phenomenon, but a structural property of inferential tasks. The study shows that, even under homogeneous instructions and with expert annotators, human interpretations diverge consistently, due to differences in prior knowledge, personal experience and cultural background. Its findings reinforce the idea that disagreement contains useful semantic information and that automatic systems should incorporate this variability rather than treating it as noise.

Other studies have explored how to make use of multiple annotations without collapsing them into a single consensus label. In [11], the authors introduce a dataset that preserves individual annotations together with rationales, showing that this approach makes it possible to train models that are more interpretable and more sensitive to subjectivity. At a more general conceptual level, [12] provide an extensive review of perspectivism in NLP, systematising different types of perspectivist datasets, existing methodological approaches, and the challenges involved in evaluating models trained with disagreement.

Finally, recent research has begun to explore how to combine models trained from different perspectives. This work [6] investigates advanced ensembling mechanisms based on the confidence level of each model. Their findings show that dynamically adjusting the contribution of each perspective increases the overall robustness of the system. Such techniques reinforce the central hypothesis of computational perspectivism: differences between annotators should not only be preserved, but can be turned into an advantage when appropriately integrated into more representative and accurate models. Along these lines, [13] demonstrate that interpretive differences are neither noisy nor arbitrary, but systematic and amenable to modelling. They further show that integrating multiple interpretations through aggregation techniques improves the final semantic representation.

Taken together, this line of work reinforces the central claim of perspectivism: annotator divergences can be turned into an advantage when they are integrated in a structured way using suitable combination methods.

3. Datasets

Despite the growing interest in annotator disagreement and perspectivism in NLP, datasets that provide multiple annotations per instance together with detailed demographic metadata are still very limited. The limited availability of such corpora hinders research on perspectivist approaches and constrains the ability to compare different methods for modelling disagreement. For this reason, creating new multi-annotator datasets, explicitly designed to capture interpretative diversity and accompanied by reliable metadata, is essential to further advance this line of research.

To evaluate the proposed approach, three datasets were used. Two of them were provided in the framework of the EXIST 2024 challenge *(sEXism Identification in Social neTworks)*: the tweet corpus (inherited from the 2023 edition) and the meme corpus, introduced in 2024 [14]. The third is a subset of SST-2 from the study *Being Right for Whose Right Reasons?* [15]. These datasets were designed under the *Learning with Disagreement* (LeWiDi) paradigm, where each instance has been annotated by multiple individuals, capturing different perspectives on the presence of sexism and on sentiments.

Table 1
Samples from the EXIST 2024 Texts dataset: textual content and labels.

Id_EXIST	Tweet	Annotators	Labels_task1	Gold_label
100673	[user] Eso es bullying. Patea a una persona cuando está en el suelo	['Annotator_385', …, 'Annotator_390']	[NO, NO, NO, NO, NO, NO]	NO
103328	[user] no se si me gusta más el sujetador o las rellenas pero sé si están bien metidas…	['Annotator_301', …, 'Annotator_306']	[NO, YES, YES, YES, YES, NO]	YES
202378	Who says you can't wear a grammar lesson? My students loved my shirt [lol]	['Annotator_467', …, 'Annotator_472']	[NO, NO, NO, NO, NO, NO]	NO

Table 2
Samples from the EXIST 2024 Texts dataset: demographic characteristics of annotators.

Id_EXIST	Annotators	Gender	Age	Ethnicities	Study_levels	Countries
100673	['Annotator_385', …]	[F, F, F, M, M, M]	['18-22', …, '18-22']	['Hispano or Latino', …, 'Hispano or Latino']	['Bachelor's degree', …, 'High school']	['Mexico', …, 'Spain']
103328	['Annotator_301', …]	[F, F, F, M, M, M]	['18-22', …, '18-22']	['White or Caucasian', …, 'Hispano o Latino']	['High school', …, 'Bachelor's']	['Spain', …, 'Portugal']
202378	['Annotator_467', …]	[F, F, M, M, M, F]	['18-22', …, '18-22']	['White or Caucasian', …, 'Hispano o Latino']	['High school', …, 'Other']	['Poland', …, 'US']

Table 3
Samples from the EXIST 2024 Memes dataset: textual content and labels.

Id_EXIST	Tweet	Annotators	Labels_task1	Gold_label
110946	DUCHATE GUARRA GRACIAS	['Annotator_211', …, 'Annotator_216']	[YES, YES, YES, YES, YES, NO]	YES
210087	LADIES PLEASE CONTAIN YOUR ORGASM	['Annotator_475', …, 'Annotator_70']	[NO, NO, NO, NO, NO, YES]	NO
111900	TIENEN SEXO ENTRE COMPAÑEROS DE TRABAJO QUE POCO SERIOS	['Annotator_421', …, 'Annotator_426']	[NO, NO, NO, YES, NO, NO]	NO

Table 4
Samples from the EXIST 2024 Memes dataset: demographic characteristics of annotators.

Id_EXIST	Annotators	Gender	Age	Ethnicities	Study_levels	Countries
110946	['Annotator_211', …]	[F, F, F, M, M, M]	['18-22', …, '23-45']	['White or Caucasian', …, 'Hispano or Latino']	['Bachelor's degree', …, 'Master's degree']	['Portugal', …, 'Chile']
210087	['Annotator_475', …]	[F, F, F, M, M, M]	['18-22', …, '46 + ']	['Hispano o Latino', …, 'White or Caucasian']	['Bachelor's degree', …, 'Bachelor's degree']	['Mexico', …, 'Spain']
111900	['Annotator_421', …]	[F, F, M, M, M, F]	['18-22', …, '23-45']	['White or Caucasian', …, 'White or Caucasian']	['Bachelor's degree', …, 'Bachelor's degree']	['Portugal', …, 'Portugal']

3.1. EXIST 2024

EXIST 2024, organised as part of CLEF, is a key benchmark for evaluating sexism detection systems, as it includes data annotated by multiple individuals and makes it possible to study how diversity of perspectives affects automatic classification.

3.1.1. Tweet dataset

The tweet corpus originates from EXIST 2023 and was reused in EXIST 2024. It contains 6920 multilingual instances (Spanish and English) in the training set and 1038 instances in the development/test set. Each tweet was annotated by six individuals for different tasks: (i) sexism detection (binary: sexist / non-sexist), (ii) author's intent (direct, reported, judgemental, unknown), and (iii) categorisation of the type of sexism. In addition to the aggregated label, the six individual labels and demographic metadata (gender, age, country, educational level and ethnicity) of the annotators are preserved.

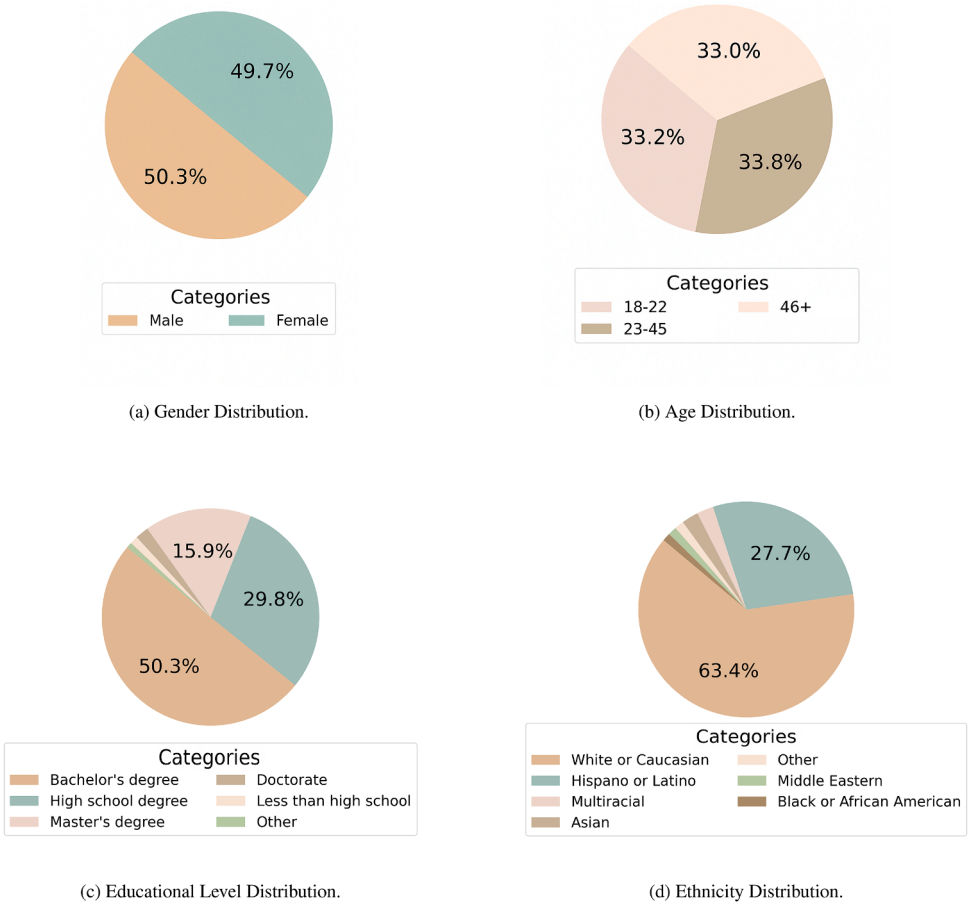
In this study, only the data provided for the binary classification task was used, focusing specifically on subjective annotator disagreement.

Although the dataset also includes a "country" variable, this feature was not taken into account in the perspectivist analysis. The geographical distribution of annotators is highly heterogeneous, with a wide variety of countries represented. Given this strong fragmentation, selecting only a subset of countries to build country-specific models would have introduced artificial biases and arbitrary decisions that are difficult to justify methodologically. Moreover, the other available demographic variables (gender, age, ethnicity and educational level) already provided a broad and representative basis for studying interpretative variability.

Table 1 and Table 2 provide an example of the corpus format, highlighting the fields available for each instance: tweet content, annotator demographics, and the labels assigned by each annotator.

The *gold_label* column was later generated by majority voting, based on the labels provided by the different annotators for each tweet. In cases of tied votes, the corresponding instance was discarded to avoid introducing ambiguity into the reference label. This step was necessary, as the original dataset did not include a single reference label, which made it impossible to coherently assess model performance during training. Having a reasonably stable aggregated label was therefore necessary to establish a reliable point of comparison.

Using this *gold_label*, a *baseline* model was trained on the full dataset, without taking individual annotator perspectives into account. This



(a) Gender Distribution.

(b) Age Distribution.

(c) Educational Level Distribution.

(d) Ethnicity Distribution.

Fig. 1. Distribution of annotators by gender, age, educational level, and ethnicity in EXIST 2024.

baseline provides a reference point for the initial reference needed to assess how much the later perspectivist models improve over the traditional single-label approach.

The dataset was annotated by 725 annotators, providing a solid basis for the perspectivist analysis. The demographic labels used, and their corresponding values, are taken directly from the original EXIST 2024 dataset, without any modifications or reclassification in this study.

The gender distribution is almost perfectly balanced, with 50.3% male and 49.7% female participants. In terms of age, annotators are evenly spread across three groups: 18-22 years (33.2%), 23-45 years (33.8%) and 46+ years (33.0%).

Regarding ethnicity, the largest group is White or Caucasian (63.4%), followed by Hispano or Latino (27.7%), with smaller proportions of Black or African American (5.8%) and multiracial and Asian minorities. Overall, this profile reflects predominantly Western participation, while still including substantial representation from Latin American groups.

Regarding educational level, more than half of the annotators have a university degree (Bachelor's, 50.3%), followed by secondary education (29.8%) and postgraduate studies (17.6% combining Master's and PhD). This distribution suggests a medium-to-high level of academic literacy, which may be associated with greater consistency in linguistic annotation.

3.1.2. Memes dataset (texts)

The meme component, introduced in EXIST 2024, extends the task to a multimodal setting by combining text and images. It was built using 250 terms (112 in English and 138 in Spanish) related to sexist expressions, which were used as queries in Google Images. After a cleaning and filtering process, the resulting corpus includes approximately 4000 memes per language. The annotations cover the same tasks as the textual corpus: binary detection, author's intent, and categorisation of the type of sexism.

As with the previous dataset, only the data provided for the binary classification task was used, and only the textual content of the memes was considered, ignoring the visual information. This choice ensures consistency with the other datasets and to focus the analysis on the linguistic component. Table 3 and Table 4 show an example from the EXIST 2024 dataset. Similarly, the *gold_label* column was generated by majority voting over the annotators' labels, in order to obtain a common reference for validation during training.

In total, this dataset was annotated by 887 annotators, 725 of whom had previously taken part in the annotation of the textual corpus, with 162 new participants joining for the annotation of this dataset. This increase in the number of annotators made it possible to maintain a stable demographic base, with distributions very similar to those of the tweet corpus: a balanced gender split (50.4% men and 49.6% women), as well as a comparable balance across the three age ranges considered (18-22, 23-45 and 46+ years).

The demographic distribution of the annotators in the meme dataset is practically equivalent to the distribution in the textual corpus, maintaining a similar balance in terms of gender, age, educational level and ethnic background. This ensures demographic consistency across datasets within the EXIST 2024 framework. As in the text corpus, the *country* variable was not used in this study, due to the high heterogeneity of origins and the lack of a solid criterion for selecting meaningful subgroups without introducing additional bias.

Table 5
Samples from the re-annotated SST-2 dataset: textual content and labels.

Text_id	Sentence	#Annotators	Annotators_IDs	Original_label	Labels
109	Kirshner and Monroe seem to be in a contest to see who can out-bad-act the other.	4	[23, 55, 66, 151]	0	[0, 0, 0, 0]
220	So young, so smart, such talent, such a wise ***.	10	[34, 47, 49, 62, 68, 98, 159, 190, 197, 206]	1	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
234	A generic family comedy unlikely to be appreciated by anyone outside the under-10 set.	5	[5, 13, 16, 65, 94]	0	[0, 1, 0, 0, 0]

Table 6
Samples from the re-annotated SST-2 dataset: demographic characteristics of annotators.

Text_id	Annotators_IDs	Genders	Ages	Age_range	Ethnicities
109	[23, …, 151]	['Male', …, 'Female']	[44, …, 50]	['Older', …, 'Older']	['White/Caucasian', …, 'Black/African American']
220	[34, …, 206]	['Male', …, 'Male']	[49, …, 38]	['Older', …, 'Middle']	['White/Caucasian', …, 'Black/African American']
234	[5, …, 94]	['Male', …, 'Female']	[39, …, 21]	['Middle', …, 'Middle']	['Black/African American', …, 'White/Caucasian']

Fig. 1 illustrates the overall demographic distribution of the annotators who took part in EXIST 2024, jointly considering both the text and meme corpora. This representation provides a global view of the diversity and balance of the main demographic variables (gender, age, ethnicity and educational level).

3.1.3. Re-annotated SST-2 dataset

This dataset is drawn from the study Being Right for Whose Right Reasons? [15], in which the authors re-annotate 263 examples from the Stanford Sentiment Treebank 2 (SST-2) with demographic metadata for the annotators. The corpus consists of sentences extracted from film reviews on the Rotten Tomatoes website (https://www.rottentomatoes.com/). Each instance in the corpus contains an English sentence associated with a binary sentiment label indicating whether the content expresses a positive (1) or negative (0) evaluation.

The annotation process consisted of a manual re-annotation of the 263 instances, where each instance was evaluated six times, once for each of the six demographic subgroups defined in the study. In total, the dataset was annotated by 156 annotators, evenly distributed across the six groups. The annotators were recruited via the Prolific platform (https://www.prolific.com/), with the aim of balancing the sample according to two criteria: ethnicity and age. The groups used to define the six subgroups were:

- Ethnicity (3 groups): Black/African American (B), Latino /Hispanic (L), and White/Caucasian (W).
- Age (2 groups): Under 36 years (Young - Y) and Over 37 years (Older - O).

The cross-product of these criteria resulted in the six annotator subgroups: BO, BY, LO, LY, WO, WY.

Although the original re-annotation process for SST-2 is organised into six groups defined by the combination of age and ethnicity, in this work we adopt a different approach in order to maintain methodological consistency with the EXIST 2024 datasets.

More concretely, the perspectives used in our study were organised according to three demographic variables-gender, age and ethnicity-replicating the same segmentation scheme applied to the other two datasets used. This adaptation made it possible to compare the behaviour of the perspectivist models across the different datasets in a homogeneous way. It also made it possible to analyse the individual impact of each demographic characteristic on sentiment classification.

Table 5 and Table 6 show an example of instances from the dataset, including the original sentences, the sentiment labels and the annotators' demographic metadata.

Unlike the EXIST 2024 datasets, where the *gold_label* column was generated specifically for this study by majority voting, the re-annotated SST-2 dataset already included an original reference label, *original_label*, provided by the corpus authors. This label indicates whether each sentence expresses a positive or negative sentiment and is used as the reference label for model validation. In addition, an *age_range* column was created, grouping annotators into three age ranges: young (18-27 years), middle (28-41 years) and older (42+ years). This division makes it possible to train three separate models according to age subgroup, maintaining the methodological coherence used for the EXIST datasets, where separate models were built for each demographic perspective. This design ensures a homogeneous approach across the three datasets analysed.

Fig. 2 shows the distribution of annotators used in this study. There is a higher proportion of White/Caucasian participants, followed by Latino/Hispanic and Black/African American annotators. The gender distribution is balanced, with very similar proportions of men and women. Similarly, shows a balanced distribution across the young (18-27), middle (28-41) and older (42+) groups, ensuring a balanced demographic profile that is suitable for perspectivist analysis.

4. Methodology

This section describes the methodology adopted, which is structured into three main stages: splitting the datasets by perspective, training independent models, and subsequently combining them using a weighted approach.

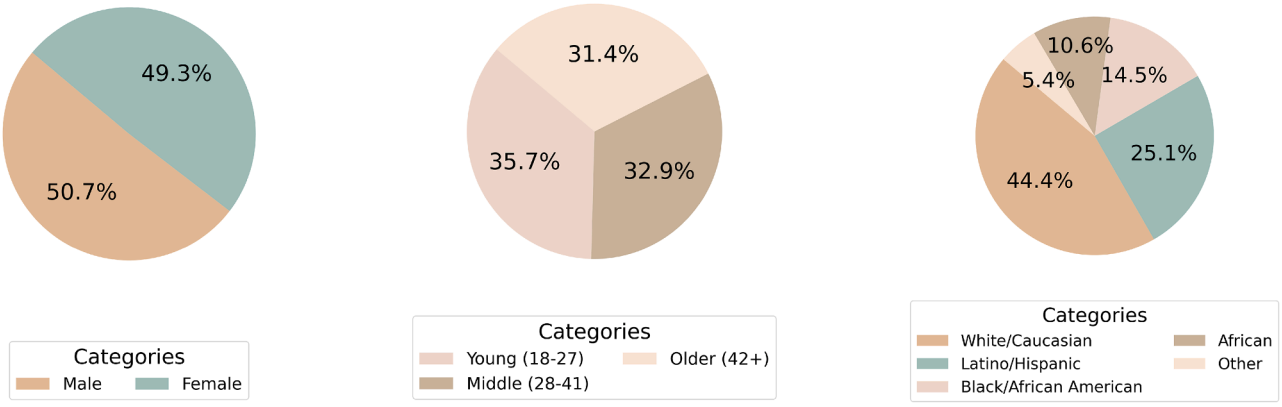
This methodology is designed to capture interpretative differences between demographic groups of annotators and to study their impact on model performance. For this purpose, the data were segmented by sociodemographic variables: gender, age, ethnicity, and educational level in the EXIST datasets, and gender, age, and ethnicity in the re-annotated SST-2 dataset. A weighted combination strategy was then applied to integrate individual predictions into a single optimised output, preserving interpretative diversity without compromising system robustness.

Before carrying out this process, a baseline model was trained using the full dataset and the majority-aggregated label: *gold_label* for the EXIST datasets (generated for this study) and *original_label* for the SST-2 set (already provided in the dataset). This initial model provides a necessary reference point. It allows us to assess whether the later perspectivist models improve performance over a traditional approach that does not incorporate demographic information.

4.1. Perspective-based modeling

4.1.1. Splitting datasets by perspective

The proposed approach is based on perspectivist modelling, designed to capture the interpretative differences that emerged between different groups of annotators. To this end, each dataset is split according to



(a) Gender Distribution. (b) Age Distribution. (c) Ethnicity Distribution.

Fig. 2. Distribution of annotators by gender, age, and ethnicity in the re-annotated SST-2 dataset.

Table 7
Number of instances per dataset.

Dataset	Train	Valid	Test
EXIST 2024 - Tweets	4851	1213	934
EXIST 2024 - Memes	2188	548	2736
Re-annotated SST-2	184	39	40

Table 8
Number of instances in each subset: *Test 1 - Weighted-calibration* test dataset and *Test 2 - Held-out* test dataset after the split.

Dataset	Test 1	Test 2
EXIST 2024 - Tweets	467	467
EXIST 2024 - Memes	1368	1368
Re-annotated SST-2	20	20

demographic metadata values: gender, age, ethnicity and education for the EXIST datasets, and gender, age and ethnicity for the re-annotated SST-2 set, which results in perspective-specific subsets. This segmentation makes it possible to train models that reflect the particular way in which each group perceives and labels the linguistic phenomena, avoiding the loss of variation introduced by consensus-based aggregated labels.

Before training, all datasets went through a basic preprocessing phase, adapted to the nature of the content. In the EXIST 2024 datasets (tweets and memes), null rows were removed and the text was cleaned of emoticons, hashtags and special characters. In addition, user mentions were replaced with the token [user] and URLs with [url], in order to preserve the linguistic structure without introducing biases associated with external information. In the case of the re-annotated SST-2 dataset, since it consists of sentences from film reviews, preprocessing was minimal and limited to the removal of null or duplicate rows.

For the dataset splits, an 80%-20% partition was applied to the official training set of the EXIST 2024 datasets (tweets and memes) for training (train) and validation (valid), respectively. In the case of the re-annotated SST-2 dataset, due to the limited number of available instances, a 70%-15%-15% split was used for training, validation and test, respectively.

In addition to the official training set in EXIST 2024 datasets, the organisers released an official validation/development split intended to be used during model development. Since this split is distributed with gold labels, we adopted it as the external evaluation reference in our experiments. Throughout the paper, the term “test set” for EXIST refers to this official validation/development set.

Building on the evaluation setting described above, to ensure a rigorous evaluation and avoid any form of overfitting during the weighted combination process, the test split used in our experiments for each dataset was carefully partitioned into two independent subsets. These are referred to as *Test 1* (*Weighting-calibration test set*) and *Test 2* (*Held-out test set*), using a stratified 50–50 split to preserve class proportions. Specifically, *Test 1* was used exclusively to estimate the individual performance of each perspectivist model; based on the F_1-score, normalised weights (and, when applicable, the decision threshold) were computed to determine each perspective’s contribution to the ensemble prediction. *Test 2* was reserved exclusively for the final held-out evaluation of the weighted ensemble, ensuring that no instances used for calibration are used for final reporting.

Test 2 was reserved exclusively as the final evaluation set for the weighted ensemble. In this way, the weights are obtained from data that are not used in the definitive test phase, ensuring that the evaluation is not carried out on data that were also used to optimise the ensemble. This separation between weight calculation and evaluation prevents the ensemble from being indirectly adapted to the test dataset, thereby preserving the objectivity of the results.

The split was performed in a stratified and balanced way (50-50), ensuring that both partitions preserved the same class proportions and the same statistical distribution as the original test set. This experimental design ensures that the weights are based on the true performance of the perspectivist models, and that the final evaluation of the ensemble is independent, fair and representative.

Table 7 shows the number of instances in the training, validation, and test datasets used in the three datasets, as well as the total number of instances considered in each case. The number of instances contained in the *Test 1* and *Test 2* splits after the division can be found in Table 8.

4.1.2. *Training independent models*

To implement the perspectivist approach, independent data subsets were generated based on each available demographic variable. Regarding the perspectives based on educational level and ethnicity within both EXIST 2024 datasets, the selection was restricted to the ‘High School´, ’Bachelor’s degree’, and ’White/Caucasian’ groups, as these were by far the most predominant. The remaining categories contained insufficient samples to train robust and comparable perspectivist models. Consequently, the analysis focused exclusively on these majority groups, ensuring model stability and methodological consistency. To avoid concluding highly underpowered models or from extreme class imbalance, we restricted perspectivist training to subgroups that provided sufficient support for reliable fine-tuning and evaluation. Concretely, we prioritised the most represented categories and omitted the remaining ones, which would otherwise yield unstable estimates and confound the comparison between perspectives. We explicitly acknowledge that this

Table 9
EXIST 2024 - Memes training dataset size before and after data augmentation.

Class	Before	After
Non Sexist (Label 0)	884	1768
Sexist (Label 1)	1304	1304

choice limits the coverage of minority subgroups, and we consider extending the analysis to additional categories (when sufficient data are available or through data-efficient modelling) as future work.

In each case, a new subgroup-specific dataset was constructed, incorporating solely the annotations provided by annotators sharing the relevant demographic characteristic. Given that the original labels represented an aggregation of all perspectives, it was necessary to generate a new column, *Gold_label_profile*, for each subset. This label was derived via majority voting using annotations from the corresponding group (e.g., exclusively those from female annotators within the female dataset). In cases of voting tie, the instance was excluded to prevent the introduction of noise during training. This procedure was systematically applied across all three datasets-EXIST 2024 (tweets and memes) and the Re-annotated SST-2 -ensuring that each perspectivist model was trained exclusively on labels derived from its specific demographic group.

To address the class imbalance observed in the EXIST 2024 Memes Texts training set, we employed a data augmentation technique based on *Contextual Word Embeddings* (CWE) [16]. Specifically, we utilized a BERT-based architecture, the *bert-base-uncased* model [17]. This method entails substituting selected tokens with semantically similar alternatives derived from the model's contextual representations, thereby increasing the lexical diversity of the corpus while preserving the semantic integrity of the original instances.

Since the imbalance affected only the negative class (non-sexist), the data augmentation process was applied exclusively to this class, increasing its number of examples to better match the size of the positive class. The positive class (sexist), which was already sufficiently represented, was left unchanged. In this way, data augmentation helped to correct the imbalance without introducing unnecessary noise into the majority class, resulting in a more stable and balanced training process between the two classes. Table 9 shows the number of instances in the dataset before and after applying this strategy.

It is worth noting that our CWE-based augmentation is designed to mitigate imbalance rather than enforce a perfectly symmetric distribution. In the EXIST 2024-Memes training split, the minority class corresponds to Non Sexist (Label 0), and augmentation is applied exclusively to that class. Because CWE generates paraphrastic variants from the original instances, the number of augmented samples is determined by the target balancing criterion and may slightly overshoot the positive class, yielding a mild inversion in the final counts (Table 9). This does not reproduce the imbalance problem in practice, as the resulting distribution remains close to balanced, and augmentation is applied only to the training data, while validation and test splits are kept unchanged to ensure an unbiased evaluation.

Tables 10–12 present the distribution of instances across demographic subgroups in the three datasets used in this study.

Once the subsets had been generated, an independent model was trained for each perspective. For the EXIST 2024 Texts and the re-annotated SST-2 datasets, we used XLM-RoBERTa-base [18], whereas for EXIST 2024 Memes we adopted mDeBERTa-v3-base [19]. This choice was motivated by the fact that XLM-RoBERTa is a widely used baseline for multilingual sentence-level classification [20], facilitating comparability with previous work on textual data. In contrast, mDeBERTa-v3 has shown strong performance on multilingual classification tasks involving shorter and noisier inputs, making it particularly suitable for meme captions. In all cases, a supervised fine-tuning process was carried out on the corresponding task: sexism detection (for the EXIST 2024 datasets) or sentiment analysis (for the re-annotated SST-2).

Table 10
Distribution of subsets in the EXIST 2024 - Texts dataset.

Demographic Variable	Subgroup	Train	Valid
Gender	Male	4851	1213
	Female	4851	1213
Age	18–22	1888	473
	23–45	1888	472
	46 +	1796	449
Educational Level	High	3966	992
	Bachelor	4500	1125
Ethnicity	White, Caucasian	4640	1161

Table 11
Distribution of subsets in the EXIST 2024 - Memes dataset.

Demographic Variable	Subgroup	Train	Valid
Gender	Male	3051	548
	Female	3037	548
Age	18–22	2192	373
	23–45	2140	382
	46 +	2046	381
Educational Level	High	2431	421
	Bachelor	2798	492
Ethnicity	White, Caucasian	2987	523

Table 12
Distribution of subsets in the Re-annotated SST-2 dataset.

Demographic Variable	Subgroup	Train	Valid
Gender	Male	189	33
	Female	190	34
Age	18–27	128	32
	28–41	175	31
	42 +	189	34
Ethnicity	White	188	33
	Latino	189	33
	Black	187	33

The hyperparameters were kept fixed to ensure comparability between models, using a *learning rate* of 3e-5, *batch size* of 16 and training for 10 epochs with *Early Stopping*. These values were chosen because they are widely used and have been shown to yield stable performance when fine-tuning Transformer-based models for text classification, particularly on social media and sentiment-related tasks [18,21,22]. In addition, the main aim of the study is not hyperparameter optimisation, but the analysis of perspectivist behaviour and the evaluation of the weighted combination method. For this reason, experimental consistency and the comparative validity of the results are prioritised over searching for model-specific configurations.

4.1.3. Metrics

Each perspectivist model was evaluated independently on its corresponding validation set, using multiple binary classification metrics [23]. Firstly, the metrics derived from the confusion matrix were computed: Precision, Recall, Accuracy and F_1-score. In addition, the Area Under the Curve (AUC) [24] was calculated as a global indicator of the model's ability to correctly discriminate between the two classes.

Among these, the metrics selected as the main performance indicators were Accuracy, F_1-score and Area Under the Curve (AUC).

The accuracy metric is the most basic performance measure in binary classification and reflects the proportion of correct predictions made by the model over the total number of instances. Although it is an intuitive and widely used metric, it may be insufficient in tasks where the classes are hard to distinguish or there is annotator disagreement. For this reason, it is complemented with more informative metrics such as F_1-score and AUC. It is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives, respectively.

The F_1-score metric provides a balanced measure of precision and recall, making it a suitable indicator for assessing overall model performance in tasks where class separation is ambiguous or labels are highly subjective. It is defined as follows:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

The AUC metric summarises the discriminative ability of the model across different decision thresholds, offering a more comprehensive view of its performance than a single-point accuracy measure. Using F_1-score and AUC together therefore makes it possible to assess both the consistency of the predictions and the model's ability to distinguish between classes in settings where human interpretations may diverge [25]. It is mathematically defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{3}$$

where TPR and FPR are the true positive rate and the false positive rate in the ROC curve[26].

In subjective tasks such as sexism detection or sentiment analysis-where class boundaries may be ambiguous and labels not always consensual-these metrics provide a more stable and representative evaluation of the model's real behaviour. The F_1-score values obtained for each perspectivist model were subsequently used in the weighted combination phase to set the weights associated with each perspective according to its individual performance. Accuracy and AUC were only used as additional reference metrics to analyse performance differences and did not play any role in the weighting scheme.

4.2. Perspective_weighted ensemble

The use of weighted model combinations (or weighted ensembling) is based on the idea that the different perspectivist classifiers capture complementary dimensions of the linguistic phenomenon, derived from demographic differences among annotators [27]. Compared to traditional uniform voting, weighted ensembling assigns greater influence to better-performing models, reducing the bias introduced by less accurate models and improving the robustness of the final system [28]. The same experimental strategy was applied consistently across all studies and datasets.

4.2.1. Weight estimation based on individual performance (Test 1 - Weighted-calibration test set)

After training each perspectivist model, its performance was evaluated independently on the first partition of the test set, referred to, as mentioned above, as *Test 1 - Weighted-calibration* test. This evaluation made it possible to measure the quality of each perspective in an objective way. The F_1-score obtained by each model on this set was then used to determine its weight in the ensemble.

The weights were normalized as (4):

$$w_i = \frac{F1_i}{\sum_{j=1}^{n} F1_j} \tag{4}$$

where w_i is the weight of the model i, $F1_i$ its individual score y n the total number of models.

Illustrative example: if three models achieve F_1 values of 0.80, 0.70, and 0.50, the resulting weights would be approximately 0.40, 0.35, and 0.25, proportionally reflecting their reliability.

4.2.2. The need for an optimal decision threshold

Since the outputs of the perspectivist models were kept in probabilistic form before applying the discrete decision, it was necessary to determine a decision threshold that optimised the conversion of probabilities into binary labels. Rather than using the standard threshold of 0.5, an optimal threshold was selected so as to maximise the ensemble's F_1-score, which is particularly important in subjective tasks where the boundary between classes is diffuse.

A total of 100 equally spaced thresholds between 0 and 1 were explored, using a step size of 0.01. For each value θ:

1. The combined probability was converted into a binary label using $p \geq \theta$.
2. The F_1-*score* was computed.
3. The threshold that maximized the F_1-*score* was selected.

The combined probability corresponds to the final value obtained by integrating the probabilistic predictions of all perspectivist models using the previously computed normalised weights. Each model M_i produces, for a given instance x, a probability $p_i(x)$ that this instance belongs to the positive class. These probabilities are not discarded or immediately converted into binary labels; instead, they are kept in their continuous form so that they can be combined. The combination is carried out by applying a weighted sum (5):

$$p_{\text{comb}}(x) = \sum_{i=1}^{n} w_i \cdot p_i(x), \tag{5}$$

where w_i is the weight assigned to model i based on its F_1-*score* on *Test 1*. This combined probability therefore reflects the relative contribution of each perspective, giving more reliable perspectives greater influence on the final prediction, while less consistent ones have a proportionally smaller impact. Once the combined probability is obtained, the optimal threshold is applied to convert it into a binary label.

Illustrative example of the combined probability calculation: Consider an instance x evaluated by eight perspectivist models. Each model M_i returns a probability $p_i(x)$ indicating the probability that the instance belongs to the positive class. Assume that the probabilities obtained are:

$$p(x) = [0.12,\ 0.30,\ 0.44,\ 0.51,\ 0.63,\ 0.72,\ 0.81,\ 0.90].$$

Likewise, the normalized weights associated with each model (calculated from its F_1-*score* on *Test 1*) are:

$$w = [0.05,\ 0.07,\ 0.10,\ 0.12,\ 0.15,\ 0.18,\ 0.15,\ 0.18].$$

The combined probability is obtained by applying the weighted sum (6):

$$p_{\text{comb}}(x) = \sum_{i=1}^{8} w_i \cdot p_i(x). \tag{6}$$

Replacing:

$$\begin{aligned} p_{\text{comb}}(x) &= 0.05(0.12) + 0.07(0.30) + 0.10(0.44) \\ &\quad + 0.12(0.51) + 0.15(0.63) + 0.18(0.72) \\ &\quad + 0.15(0.81) + 0.18(0.90) \\ &= 0.006 + 0.021 + 0.044 \\ &\quad + 0.061 + 0.0945 + 0.1296 \\ &\quad + 0.1215 + 0.162 \\ &= 0.6396. \end{aligned}$$

In this example, the final combined probability is:

$$p_{\text{comb}}(x) \approx 0.64.$$

This value summarizes the proportional contribution of all perspectives: The more reliable models (i.e., those with higher F_1-*score*) have

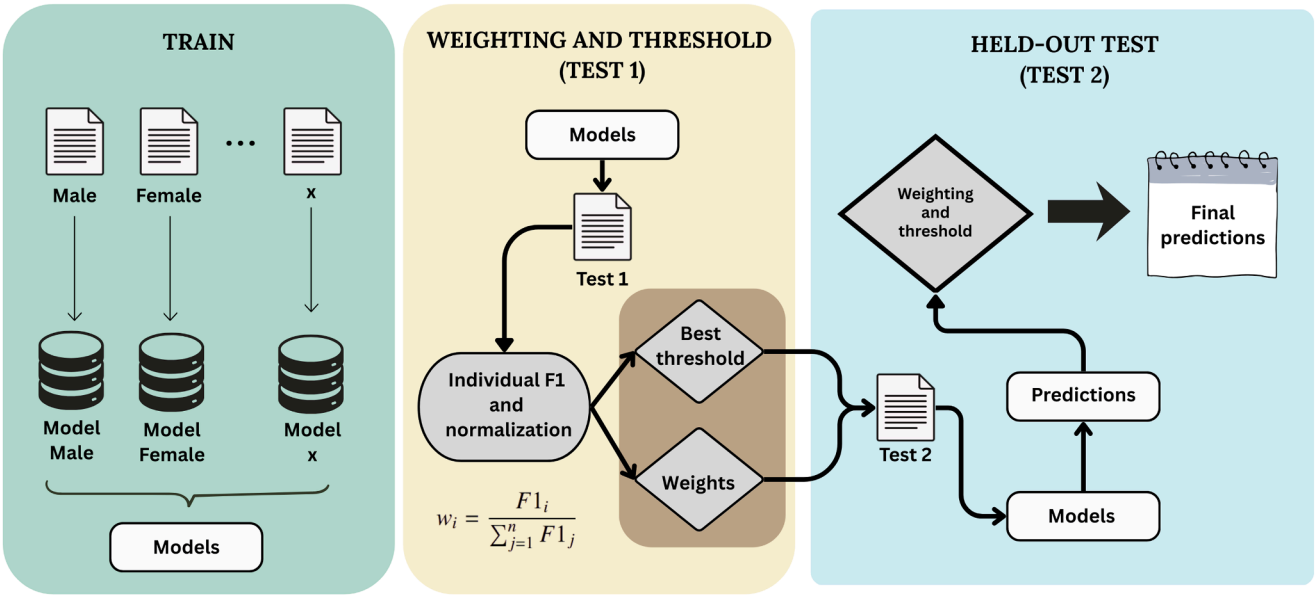


Fig. 3. Overview of the weighted combination process for perspectivist models.

Table 13
Final probabilities obtained.

Instance	Combined Probability	Label
1	0.78	1
2	0.62	0
3	0.49	1
4	0.44	1
5	0.31	0

a greater influence on the result than those with lower performance. Subsequently, this combined probability is compared with the optimal threshold θ to obtain the final binary label on *Test 2*.

Illustrative example for selecting the optimal threshold: We consider a simplified example with a set of 5 instances that belong to *Test 1*. Suppose that, after combining the probabilities of the perspectivist models using the normalized weights, the following final probabilities final labels are obtained (see Table 13):

Based on these probabilities, it can be observed that a standard threshold of 0.5 would lead to poorly balanced predictions [29]. To illustrate this, three representative thresholds are compared:

- **threshold** $\theta = 0.50$:
  - Predictions: [1, 1, 0, 0, 0]
  - TP = 1, FP = 1, FN = 2
  - $F_1 = 0.40$
- **threshold** $\theta = 0.45$:
  - Predictions: [1, 1, 1, 0, 0]
  - TP = 2, FP = 1, FN = 1
  - $F_1 = 0.57$
- **threshold** $\theta = 0.40$:
  - Predictions: [1, 1, 1, 1, 0]
  - TP = 3, FP = 1, FN = 0
  - $F_1 = 0.75$

In this case, the threshold $\theta = 0.40$ yields the highest F_1-*score*, as it correctly recovers the three positive cases without excessively increasing the number of false positives. This example illustrates that, in subjective tasks, explicit threshold optimization can substantially improve ensemble performance, avoiding the limitations of the fixed 0.5 threshold.

This procedure ensures that the ensemble is specifically adapted to the actual probability distribution generated by its individual models.

4.2.3. *Final evaluation using Test 2 - Held-out test set*

Once the weights and threshold were estimated (as explained previously, using only *Test 1*), the weighted combination was evaluated on the second independent partition of the test set, referred to as *Test 2*. This experimental design prevents any contamination of the test set and ensures a strictly objective evaluation.

In this phase:

1. Each model generated probabilities for the instances in *Test 2*.
2. The probabilities were combined using:

$$p_{\text{ensemble}} = \sum_{i=1}^{n} w_i \cdot p_i$$

3. The optimal threshold was applied to obtain the final label.
4. The global metrics of the ensemble were computed.

Fig. 3 visually summarizes the procedure, presenting a schematic overview of each stage of the process and the relationships among them.

Taken together, this methodology ensures that (i) the weights reflect the actual performance of each perspective, (ii) the threshold is adapted to the ensemble's probability distributions, and (iii) the final evaluation is carried out entirely independently. This provides a rigorous, reproducible, and overfitting-free assessment of the perspectivist system.

5. Experimental setup

All experiments were conducted in the Google Colab Pro+ execution environment (https://colab.google/), which provides access to high-performance cloud computing resources, ensuring a balance between computational capacity and scalability. A NVIDIA A100 GPU with 40 GB of VRAM was used, accompanied by a 12-core Intel Xeon CPU and 52 GB of RAM. This environment enabled efficient and parallel training of multiple perspectivist models, optimizing training and validation times.

The development environment was based on Python 3.10, using the PyTorch library (v2.1) and Hugging Face Transformers (v4.37) for implementation, fine-tuning, and management of the pretrained XLM-RoBERTa-base and DeBERTa-v3-base models. Preprocessing and data manipulation were carried out with Pandas (v2.2) and NumPy (v1.26), while evaluation metrics were computed using scikit-learn (v1.3). To ensure reproducibility, random seeds were fixed across all used libraries,

Table 14
EXIST 2024 (Texts) results on *Test 1 - Weighted-calibration* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	0.8114	0.8300	0.8616
Male	0.8244	0.8242	0.8991
Female	0.8266	0.8262	0.8881
18–22	0.8287	0.8286	0.8841
23–45	0.8158	0.8156	0.8972
46+	0.7923	0.7919	0.8683
Bachelor	0.8330	0.8328	**0.9099**
High	0.8287	0.8282	0.8753
White Caucasian	**0.8373**	**0.8371**	0.9002

Table 15
EXIST 2024 (Memes) Results on *Test 1 - Weighted-calibration* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	**0.8997**	**0.8632**	**0.9611**
Male	0.8355	0.8320	0.8823
Female	0.8677	0.8611	0.9020
18–22	0.8048	0.7987	0.8554
23–45	0.7822	0.7759	0.8398
46+	0.7588	0.7448	0.8250
Bachelor	0.8041	0.7955	0.8692
High	0.7968	0.7859	0.8169
White Caucasian	0.8114	0.7970	0.8616

Table 16
Re-annotated SST-2 results on *Test 1 - Weighted-calibration* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	0.9000	0.8889	0.9192
Male	0.8500	0.8400	0.8586
Female	0.8000	0.7917	0.9192
18–27	0.8000	0.7917	0.9293
28–41	0.7500	0.7151	0.8384
42+	0.9000	0.8958	0.8586
White	0.8500	0.8465	0.8990
Latino	0.8500	0.8465	0.9293
Black	**0.9500**	**0.9488**	**0.9596**

Table 17
EXIST 2024 Texts results on *Test 2 - Held-out* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	0.8223	0.8218	0.8924
Male	0.7901	0.7901	0.8868
Female	0.8030	0.8024	0.8741
18–22	0.8051	0.8051	0.8717
23–45	0.8137	0.8137	0.8614
46+	0.7645	0.7630	0.8593
Bachelor	0.8201	0.8201	0.8740
High	0.7837	0.7834	0.8579
White Caucasian	0.8223	0.8220	0.8838
Weighted Ensemble	**0.8394**	**0.8393**	**0.8991**

Table 18
EXIST 2024 Memes results on *Test 2 - Held-out* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	0.8414	0.8265	0.9189
Male	0.8333	0.8309	0.8945
Female	0.8830	0.8773	0.9160
18–22	0.8414	0.8363	0.9028
23–45	0.7924	0.7856	0.8508
46+	0.7961	0.7837	0.8524
Bachelor	0.8326	0.8270	0.8892
High	0.8070	0.7971	0.8345
White Caucasian	0.8034	0.7866	0.8626
Weighted Ensemble	**0.9152**	**0.9120**	**0.9762**

and the final results of each experiment were logged. Model management and result organization were handled through a storage structure organized by demographic subgroups and tasks, allowing detailed tracking of each perspective's performance.

To ensure reproducibility, random seeds were fixed across all used libraries, and the final results of each experiment were logged. Concretely, a single global seed (`SEED = [42]`) was used consistently for dataset shuffling and split generation, and model training in Python/NumPy/PyTorch. Unless otherwise stated, all results reported in this paper correspond to a single training run per model configuration. While minor GPU-related non-determinism may still occur, fixing the seed substantially reduces run-to-run variability and facilitates faithful replication of the pipeline.

The complete source code used to train and evaluate the models are available in an online repository.[1]

6. Results

This section reports the results obtained by applying the described methodology to the three datasets considered: EXIST 2024 - Texts, EXIST 2024 - Memes, and the Re-annotated SST-2. Before training the perspectivist models, a baseline model was built on the full dataset without distinguishing among demographic perspectives, providing a reference point for evaluating the real utility of the perspectivist approach.

Before analysing the ensemble performance, the results obtained by each perspectivist model are examined on *Test 1*, the subset specifically used for weight estimation. This set allows objectively evaluate the individual performance of each demographic perspective without affecting the final system evaluation. Tables 14, 15 and 16 report the metrics obtained by the models trained for each group; their F_1-score values were then used to compute the normalised weights in the weighted combination.

Once the weights and the optimal threshold had been estimated from the *Test 1* results, the final performance of the weighted ensemble was evaluated on the independent *Test 2* subset, which was reserved exclusively for final evaluation. This set makes it possible to objectively measure the ability of the combined system to integrate the different demographic perspectives and to improve, where possible, performance with respect to both the baseline model and the individual models.

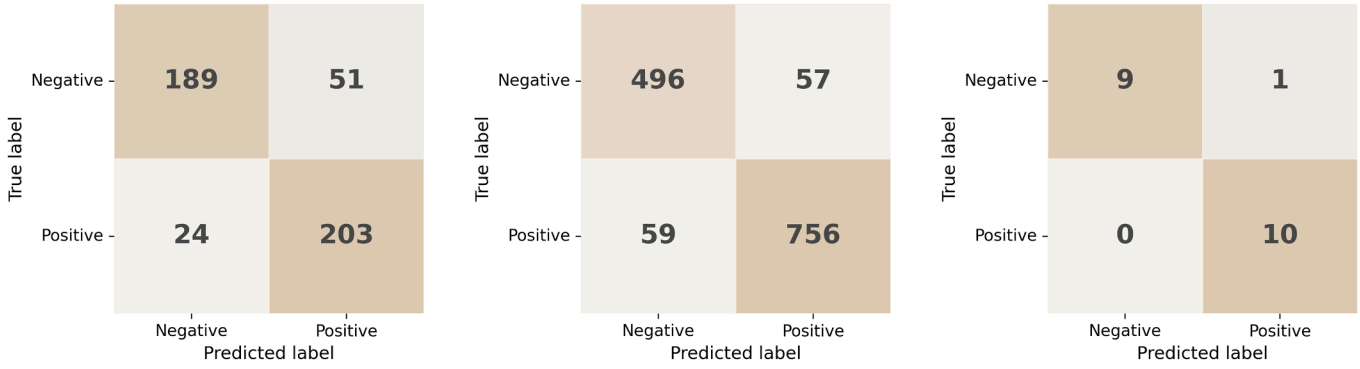
Table 17–19 shows individual metrics for each perspective model using *Test 2* and the final metrics achieved by the ensemble on each dataset, reflecting the real impact of the weighted combination on data that were not used during the calibration phase.

The proposed weighted combination technique shows consistent improvements across most datasets, although with different degrees of impact across tasks. In the case of EXIST 2024 Texts, the weighted ensemble increases the F_1-score from 0.8218 to 0.8393, AUC from 0.8924 to 0.8991 and Accuracy from 0.8223 to 0.8394. This result suggests that integrating demographic perspectives leads to a more balanced classification without significantly altering the model's overall behavior. For EXIST 2024 Memes, the proposed approach shows the most substantial improvement, reaching an F_1-score of 0.9120 and an AUC of 0.9762, compared with 0.8265 and 0.9189 for the baseline model. The simultaneous gains across all metrics support the high usefulness of the weighted method.

For the Re-annotated SST-2 dataset, the ensemble achieves an F_1-score of 0.9499, outperforming the baseline value of 0.90. Accuracy and AUC values reach values of 0.95 and 0.92, respectively. This behaviour suggests that the technique is particularly effective at improving classification quality in borderline or ambiguous cases.

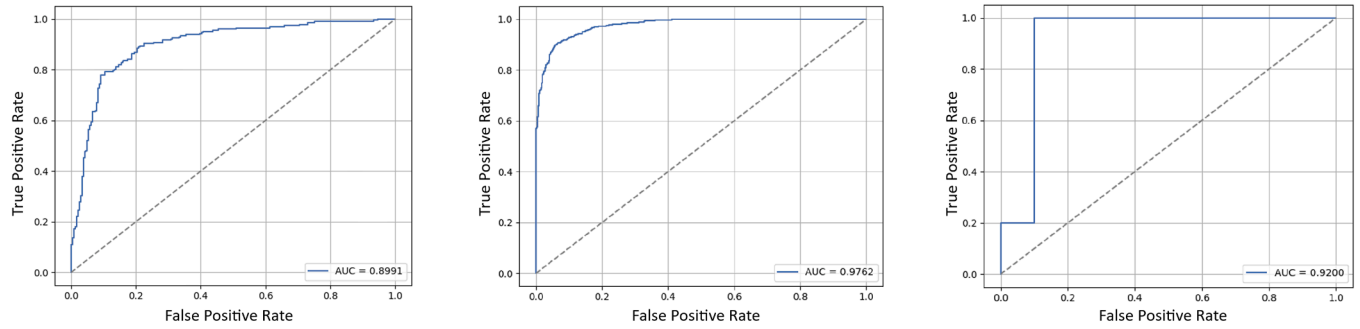
The results confirm that a weighted combination based on the individual performance of the models can enhance the discriminative capacity of the system without compromising its stability. Although the improvements are moderate, they demonstrate the potential of the perspectivist approach to capture interpretative differences across demographic groups of annotators.

[1] https://github.com/lauritavr94/perspectivist-nlp-study



(a) EXIST 2024 - Texts (b) EXIST 2024 - Memes (c) Re-annotated SST2

Fig. 4. Confusion matrices - final evaluation with *Test 2* for the weighted ensemble.



(a) EXIST 2024 - Texts (b) EXIST 2024 - Memes (c) Re-annotated SST2

Fig. 5. ROC curves - final evaluation with *Test 2* for the weighted ensemble.

Table 19
Re-annotated SST-2 results on *Test 2 - Held-out* test.

Model	Accuracy	Macro F_1-Score	AUC
Baseline	0.9000	0.9000	0.9100
Male	0.9500	0.9499	**0.9300**
Female	0.9000	0.9000	0.8900
18–27	0.8500	0.8496	0.8800
28–41	0.8500	0.8465	0.9200
42+	0.8500	0.8496	0.8500
White	0.8500	0.8496	0.9000
Latino	0.9000	0.9000	0.9800
Black	0.8000	0.7980	0.8100
Weighted Ensemble	**0.9500**	**0.9499**	0.9200

6.1. Confusion matrix and AUC curves

This subsection analyzes the confusion matrix and AUC curves obtained after applying the ensemble technique to the *Test 2* set, enabling a more precise evaluation of the system's performance in terms of errors, sensitivity, and class balance.

As shown in Fig. 4, the confusion matrix highlights the effectiveness of the models across all three datasets. For both the EXIST 2024 Texts corpus and the EXIST 2024 Memes corpus, the classifier displays a high discriminative capacity, with a clear predominance along the main diagonal (high True Positive and True Negative rates) and a comparatively low number of errors. This suggests robust generalisation, with the model successfully adapting to both textual data and memes data. The performance on the Re-annotated SST-2 dataset is particularly notable, where the model achieves nearly perfect precision: it commits zero False Positives (0) and only one False Negative (1) across the entire test set. Although this result is obtained on a smaller dataset, it underscores exceptional precision and high reliability in identifying the positive class.

Similarly, the AUC curves obtained for the three datasets (see Fig. 5) consistently demonstrate the discriminative capacity of the weighted ensemble. In all three cases, the curves lie clearly above the non-discrimination diagonal, confirming that the system is able to accurately distinguish between positive and negative instances, even under complex or subjective distributions.

The highest AUC value is obtained for EXIST 2024 - Memes (AUC = 0.9762), where the ensemble achieves an almost optimal behaviour, in line with the lower ambiguity of the phenomenon and the greater internal regularity of the dataset. In the case of the re-annotated SST-2 set, the curve also shows strong performance (AUC = 0.9200), indicating that the method is robust even when linguistic signals are combined with visual elements and a higher degree of subjectivity. Finally, the EXIST 2024 Texts dataset reaches a competitive AUC (0.8991), showing that, despite the natural variability present in this type of social content, the weighted ensemble maintains a high discriminative capacity.

6.2. Qualitative analysis of the weighted ensemble

This section presents several representative cases in which the individual models show limited performance, while the weighted ensemble corrects the final prediction. These examples illustrate how combining demographic perspectives offers complementary information which, when integrated, improves the system's robustness on conflicting or ambiguous instances. This qualitative analysis highlights the role of the ensemble as a mechanism for mitigating errors and integrating evidence from heterogeneous models.

This analysis confirms that the weighted ensemble effectively mitigates errors arising from variability across demographic models. In the two EXIST datasets, where interpretations may diverge due to irony or

Table 20
Qualitative study examples illustrating the effectiveness of the ensemble, where the column "#Models Wrong" indicates the number of models that misclassified the instance, and the columns "Male, Female, …" show the predictions of each model according to its individual perspective.

Dataset	Text	Gold_label	Weighted Ensemble	#Models Wrong	Male	Female	18_22	23_45	46+	Bachelor	High	White
EXIST 2024	[user] I dont care what gender race religionsexual orientation or background someone is if youre a good person then youre good with me	1	1	6	0.4902	0.4011	0.3835	0.9993	0.8791	0.1023	0.1351	0.3805
	[user] It just shows how money rules the world and morality is optional	1	1	6	0.1687	0.3969	0.9368	0.9997	0.4625	0.1802	0.0437	0.0999
	Calls to clean up Parliaments laddish culture while also making the British workforce more economically productive are contradictory	1	1	6	0.3753	0.2545	0.9872	0.9978	0.4946	0.1572	0.1978	0.4264

Dataset	Text	Gold_label	Weighted Ensemble	#Models Wrong	Male	Female	Young	Middle	Older	White	Latino	Black
SST-2	Narc is a no-bull throwback to 1970s action filmmaking.	1	1	4	0.9275	0.9801	0.3187	0.6329	0.0219	0.1101	0.9953	0.8163
	Pacino and Williams seem to keep upping the ante with each movie.	1	1	4	0.9864	0.0049	0.4304	0.6799	0.0255	0.9150	0.9975	0.8391

Table 21
Paired bootstrap analysis on *Test 2*: absolute Macro F_1-score differences (ΔF_1) between the weighted ensemble and the baseline with 95% confidence intervals estimated via non-parametric bootstrap ($B = 5000$).

Dataset (Test 2)	n	F_1 Baseline	F_1 Ensemble	ΔF_1 (95% CI)
EXIST Tweets	467	0.813	0.844	+0.031 [-0.001, +0.065]
EXIST Memes	1368	0.754	0.928	+0.174 [+0.151, +0.199]
SST-2	20	0.900	0.952	+0.052 [0.000, +0.194]

cultural references, cases occur in which 6 out of 8 individual models fail simultaneously, yet the ensemble produces the correct prediction. As shown in Table 20, this behaviour indicates that the weighted combination attenuates subgroup-specific biases and captures more stable signals at a global level.

In the case of SST-2, although discrepancies between models are less frequent due to the more homogeneous nature of the corpus, there are still examples in which the ensemble corrects errors made by three or four individual models when dealing with subjective or semantically ambiguous texts.

Across all three datasets, the weighted ensemble demonstrates that aggregating demographic models enhances robustness against errors caused by subgroup-specific biases, differing interpretations and cultural nuances. The analysed cases show that:

- Individual models capture different facets of language,
- Their errors do not always overlap,
- The weighted combination acts as a correction and stabilization mechanism.

This supports the use of weighted ensembles as an effective strategy for mitigating variability and bias, especially when models are trained from distinct demographic perspectives.

6.3. Statistical analysis

In addition to reporting point estimates of standard classification metrics, we include a statistical robustness analysis to assess whether the observed gains of the weighted perspectivist ensemble over the baseline are stable under sampling variability. We report 95% confidence intervals (CI) [30] for the paired difference in Macro F_1-score between the weighted ensemble and the baseline on the held-out evaluation split (*Test 2*). Confidence intervals are computed via non-parametric bootstrap resampling of test instances with replacement ($B = 5000$).

As we can see in Table 21, the weighted ensemble achieves higher point estimates than the baseline, but the strength of the evidence varies with dataset size and task characteristics. On EXIST Memes, the improvement is large ($\Delta F_1 = +0.174$) and the 95% confidence interval does not include zero, providing strong support that the gain is robust under instance resampling. This suggests that, in settings with higher subjectivity and/or disagreement signals, combining perspectivist models with performance-based weights yields consistent and substantial benefits.

On EXIST Tweets, the observed gain is smaller ($\Delta F_1 = +0.031$) and its confidence interval slightly overlaps zero, indicating that the improvement is compatible with a modest positive effect but is less conclusive under bootstrap uncertainty on this split. For the re-annotated SST-2, the point estimate is positive ($\Delta F_1 = +0.052$), yet the interval is wide due to the small test size ($n = 20$), which limits statistical power. Overall, these results align with our main findings: weighted perspectivist ensembling tends to be most beneficial in tasks or datasets with stronger subjectivity and disagreement, while smaller or less ambiguous settings may require larger evaluation samples (or additional runs) to establish improvements more conclusively.

7. Discussion

7.1. Critical analysis of the results

Results across the three tasks show that the weighted combination strategy consistently helps to improve the overall performance of the perspectivist models. The analysis of the assigned weights reveals that not all perspectives contribute equally to the classification process: models associated with larger groups and more internally consistent annotations-such as annotators with a *Bachelor*'s degree or those belonging to the *White_Caucasian* category-tend to receive higher weights, reflecting greater reliability in their predictions.

On the other hand, minority groups or those with greater dispersion in their judgements (for example, older annotators or certain less represented ethnic groups) contribute less to the ensemble, which suggests that the technique achieves a balance between diversity and accuracy. The weighted perspectivist approach performs particularly well in tasks characterised by high subjectivity and disagreement.

In such cases, combining perspectives makes it possible to capture interpretative nuances that a single model, or one trained on consensus labels, would typically fail to reflect. However, in more homogeneous settings with low variability in the annotations, such as the re-annotated SST-2 dataset, the gain, although positive, is more moderate, indicating that the benefit of the method depends directly on the degree of disagreement among annotators (being annotators > 2). These findings reinforce the hypothesis that disagreement should not be regarded as noise, but as a valuable source of information.

The weighted integration of perspectives makes it possible to organise and use this diversity in a systematic way, preserving interpretative differences without compromising the stability of the final model. This balance is one of the main contributions of the approach, as it mediates

between consensus-oriented annotation and a plural representation of meaning, echoing the call to incorporate disagreement into language model evaluation [3].

7.2. Advantages and limitations of the approach

The weighted perspectivist approach offers several advantages over traditional label aggregation methods. First, it preserves the interpretative diversity of annotators instead of reducing it to a single consensual "truth", which is particularly relevant in highly subjective tasks such as sexism detection or sentiment analysis. In addition, incorporating weights derived from the individual performance of each perspective provides an empirical and adaptive mechanism for integrating the different views, thereby improving the robustness of the final model. This strategy offers a flexible framework that can be applied to different domains and tasks, provided that demographic metadata about the annotators are available.

However, the method also has important limitations. One of the main ones is its reliance on annotations enriched with detailed demographic metadata, which are still relatively uncommon in many current NLP corpora. The absence of such information makes it impossible to segment the data into perspectives and therefore the direct application of the approach. Moreover, even when metadata exist, they may be incomplete or subject to privacy and ethical constraints that restrict their use, which can further limit the applicability and generalisability of perspectivist modelling in real-world settings. In addition, the need to train multiple independent models for each subgroup significantly increases the computational load and storage requirements, which may limit feasibility in low-resource contexts, production environments, or large-scale datasets.

Moreover, although the weighted combination yields consistent improvements, these are not uniform across all cases: the results suggest that the effectiveness of the method depends on the degree of disagreement and on the balance between demographic groups. In settings with highly homogeneous annotations or very limited representation of certain groups, the benefit may be minimal. Finally, while we report standard metrics (Accuracy, F_1-score, and AUC) and adopt a held-out calibration/evaluation protocol, our current study does not systematically quantify variability across multiple random seeds, nor does it provide formal statistical significance tests or confidence intervals for all comparisons. Incorporating multi-seed experiments and confidence intervals/significance testing would strengthen the robustness of the empirical claims.

Even so, the approach offers a promising path towards models that are more inclusive and sensitive to interpretative diversity, contributing to a deeper understanding of language and human judgement in linguistic processing tasks.

8. Conclusions and future works

This paper presents a perspectivist approach that combines models trained from different demographic perspectives, leveraging interpretive diversity for subjective natural language processing tasks. Unlike consensus-based aggregation that removes disagreement, the proposed approach demonstrates that it is possible to preserve annotator differences can provide useful information for improving model robustness and fairness.

The results across the three datasets-EXIST 2024 (texts and memes) and the Re-annotated SST-2-show that models trained with different perspectives can capture interpretive nuances that a single model cannot. Likewise, the weighted model combination proved an effective strategy for integrating these perspectives, offering consistent improvements over simple voting or uniform ensemble approaches, particularly in tasks with greater subjectivity. The analyses further indicate that annotator diversity and demographic balance shape overall effectiveness, highlighting the need for heterogeneous and representative annotations. Overall, the results support the applicability of the methodology across different types of datasets and disagreement-driven tasks, providing an adaptable framework that can be extended to multilingual or semantically complex contexts.

As future work, we plan to extend the approach to other types of subjective classification tasks, such as irony or emotion detection, where interpretative variability is often driven by pragmatic cues and rhetorical phenomena. In particular, recent work has highlighted the potential of perspectivist analyses for irony by relating disagreement to rhetorical figures and interpretation strategies [31]. Building on this line, we aim to evaluate whether weighted perspectivist ensembling can better capture such pragmatic ambiguity than consensus-based aggregation.

We also plan to complement our current evaluation with *soft metrics* that explicitly account for disagreement, which may provide a more faithful assessment than hard-label scores in highly subjective settings [32]. In addition, we intend to apply the methodology to larger-scale language models and to explore more sophisticated weighting mechanisms that incorporate uncertainty estimates or annotator-specific confidence signals. Finally, developing larger and more balanced perspectivist datasets with reliable demographic metadata will be essential for further advancing perspectivism-aware NLP models.

CRediT authorship contribution statement

Laura Vázquez Ramos: Writing - review & editing, Writing - original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization; **Jacinto Mata Vázquez:** Writing - review & editing, Writing - original draft, Validation, Supervision, Methodology, Investigation, Conceptualization; **Victoria Pachón Álvarez:** Writing - review & editing, Writing - original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

Data availability

The datasets used in this study are publicly available and are referenced in the manuscript

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Mansoor, E. French, M. Ali, Demographic diversity, processes and outcomes: an integrated multilevel framework, Manag. Res. Rev. 43 (5) (2020) 521–543. https://doi.org/10.1108/MRR-10-2018-0410

[2] S. Frenda, A. Pedrani, V. Basile, S.M. Lo, A.T. Cignarella, R. Panizzon, C. Marco, B. Scarlini, V. Patti, C. Bosco, D. Bernardi, EPIC: Multi-Perspective annotation of a corpus of irony, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61St Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13844–13857. https://aclanthology.org/2023.acl-long.774/. https://doi.org/10.18653/v1/2023.acl-long.774

[3] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to consider disagreement in evaluation, in: K. Church, M. Liberman, V. Kordoni (Eds.), Proceedings of the 1St Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. https://aclanthology.org/2021.bppf-1.3/. https://doi.org/10.18653/v1/2021.bppf-1.3

[4] E. Leonardelli, S. Casola, S. Peng, G. Rizzi, V. Basile, E. Fersini, D. Frassinelli, H. Jang, M. Pavlovic, B. Plank, et al., Lewidi-2025 at NLPerspectives: the third edition of the learning with disagreements shared task, in: Proceedings of the the 4Th Workshop on Perspectivist Approaches to NLP, 2025, pp. 182–195.

[5] A. Mostafazadeh Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: looking beyond the majority vote in subjective annotations, Trans. Assoc. Comput. Linguistics 10 (2022) 92–110. https://aclanthology.org/2022.tacl-1.6/. https://doi.org/10.1162/tacl_a_00449

[6] S. Casola, S.M. Lo, V. Basile, S. Frenda, A.T. Cignarella, V. Patti, C. Bosco, Confidence-based ensembling of perspective-aware models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3496–3507. https://aclanthology.org/2023.emnlp-main.212/. https://doi.org/10.18653/v1/2023.emnlp-main.212

[7] Á. Carrillo-Casado, J. Román-Pásaro, J. Mata-Vázquez, V. Pachón-Álvarez, et al., I2C-UHU At EXIST 2024: transformer-based detection of sexism and source intention in memes using a learning with disagreement approach, Working Notes of CLEF, 3740, 978–992 (2024).
[8] S.L. Blodgett, S. Barocas, H. Daumé, III, H. Wallach, Language (technology) is power: a critical survey of "Bias" in NLP, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58Th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. https://aclanthology.org/2020.acl-main.485/. https://doi.org/10.18653/v1/2020.acl-main.485
[9] D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing, Lang. Linguist. Compass 15 (8) (2021) e12432. https://doi.org/10.1111/lnc3.12432
[10] E. Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, Trans. Assoc. Comput. Linguistics 7 (2019) 677–694. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00293/1923015/tacl_a_00293.pdf. https://doi.org/10.1162/tacl_a_00293
[11] B. Mathew, P. Saha, S.M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection, 2022. https://arxiv.org/abs/2012.10289. https://doi.org/10.1609/aaai.v35i17.17745
[12] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A.T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: a survey, Lang. Resour. Eval. 59 (2) (2025) 1719–1746. https://doi.org/10.1007/s10579-024-09766-4
[13] Z. Žabokrtský, D. Zeman, M. Ševčíková, Sentence meaning representations across languages: what can we learn from existing frameworks?, Comput. Linguistics 46 (3) (2020) 605–665. https://aclanthology.org/2020.cl-3.3/. https://doi.org/10.1162/coli_a_00385
[14] 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 tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117. https://doi.org/10.1007/978-3-031-71908-0_5
[15] T.S. Thorn Jakobsen, L. Cabello, A. Søgaard, Being right for whose right reasons?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61St Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1033–1054. https://aclanthology.org/2023.acl-long.59/. https://doi.org/10.18653/v1/2023.acl-long.59
[16] E. Sezerer, S. Tekir, A Survey On Neural Word Embeddings, 2021. https://doi.org/10.48550/arXiv.2110.01804
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. https://aclanthology.org/N19-1423. https://doi.org/10.18653/v1/N19-1423
[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58Th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. https://aclanthology.org/2020.acl-main.747/. https://doi.org/10.18653/v1/2020.acl-main.747
[19] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint, arXiv:2111.09543 (2021).
[20] X. Ou, H. Li, YNU@ Dravidian-Codemix-FIRE2020: XLM-Roberta for multi-language sentiment analysis, in: FIRE (Working Notes), 2020, pp. 560–565.
[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint , arXiv:1907.11692 (2019).
[23] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv preprint, arXiv:2008.05756 (2020). https://doi.org/10.48550/arXiv.2008.05756
[24] J.V. Carter, J. Pan, S.N. Rai, S. Galandiuk, ROC-Ing along: evaluation and interpretation of receiver operating characteristic curves, Surgery 159 (6) (2016) 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029
[25] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 6. https://doi.org/10.1186/s12864-019-6413-7
[26] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2
[27] M. Hasnain, I. Ghani, A. Ali, et al., Ensemble learning models for classification and selection of web services: a review, Comput. Syst. Sci. Eng. 40 (1) (2022). https://doi.org/10.32604/csse.2022.018300
[28] I.E. Marouf, S. Roy, E. Tartaglione, S. Lathuilière, Weighted ensemble models are strong continual learners, in: European Conference on Computer Vision, Springer, 2024, pp. 306–324.
[29] D.J. Hand, Assessing the performance of classification methods, Int. Stat. Rev. 80 (3) (2012) 400–414. https://doi.org/10.1111/j.1751-5823.2012.00183.x
[30] C. Thiele, G. Hirschfeld, Confidence intervals and sample size planning for optimal cutpoints, PLoS ONE 18 (1) (2023) e0279693.
[31] P.F. Balestrucci, M. Oliverio, E. Chierchiello, E. Di Palma, L. Anselma, V. Basile, C. Bosco, A. Mazzei, V. Patti, Towards a perspectivist understanding of irony through rhetorical figures, in: Proceedings of the 4Th Workshop on Perspectivist Approaches to NLP, 2025, pp. 27–36.
[32] G. Rizzi, E. Leonardelli, M. Poesio, A. Uma, M. Pavlovic, S. Paun, P. Rosso, E. Fersini, Soft metrics for evaluation with disagreements: an assessment, in: Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives), Co-located with LREC-COLING 2024, 2024.