Are you sure? Measuring models bias in content moderation through uncertainty

Alessandra Urbinati[1], Mirko Lai[4,3], Simona Frenda[2,3], Marco Antonio Stranisci[5,3]

[1]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, USA, [2]Heriot-Watt University, Edinburgh, Scotland, [3]aequa-tech, Torino, Italy, [4]Università del Piemonte Orientale, Alessandria, Italy, [5]Università degli Studi di Torino, Torino, Italy

Correspondence: marcoantonio.stranisci@unito.it

Abstract

Automatic content moderation is crucial to ensuring safety in social media. Language Model-based classifiers are being increasingly adopted for this task, but it has been shown that they perpetuate racial and social biases. Even if several resources and benchmark corpora have been developed to challenge this issue, measuring the fairness of models in content moderation remains an open issue. In this work, we present an unsupervised approach that benchmarks models on the basis of their uncertainty in classifying messages annotated by people belonging to vulnerable groups. We use uncertainty, computed by means of the conformal prediction technique, as a proxy to analyze the bias of 11 models against women and non-white annotators and observe to what extent it diverges from metrics based on performance, such as the F_1 score. The results show that some pre-trained models predict with high accuracy the labels coming from minority groups, even if the confidence in their prediction is low. Therefore, by measuring the confidence of models, we are able to see which groups of annotators are better represented in pre-trained models and lead the debiasing process of these models before their effective use.[1]

1 Introduction

Language models (LMs), including large language models (LLMs), have been widely used in real-world applications since their introduction due to their ability to generate and understand human-like text and have been adopted as street-level algorithms (Alkhatib and Bernstein, 2019): technologies that are implemented to enforce the rules of platforms based on user-generated content. In such a sense, these technologies play the role of bureaucrats, as they provide an interpretation of community guidelines and decide accordingly which users' behaviors must be removed and which not. Street-level algorithms are widely present in social media (Jiang et al., 2020), but their adoption is not limited to this context. Toxicity classifiers such as the Perspective API (Hosseini et al., 2017) are used to filter out unwanted texts from pretraining data, playing a crucial role in the development of fair models. Dignum (2023) highlights the over-reliance on data of these technologies, leaving the power of what the models know and how to address the problems to those who produce and maintain the data. The societal cost of doing that is high, and it has been demonstrated that such technologies tend to perpetuate social biases against vulnerable minorities, whose opinion is less represented or excluded in data (Kalluri, 2020).

In this paper, we investigate the presence of biases towards underrepresented groups in hate speech detection, using two datasets, Social Bias Inference Corpus (Sap et al., 2020) and CREHate (Lee et al., 2024), and across two dimensions: gender and ethnicity. As *bias*, we refer to systematic discrimination against a disadvantaged group of people (Friedman and Nissenbaum, 1996), who will suffer from *representational harms* because systems tend to fail to recognize their existence (Wang et al., 2022). We think that measuring the uncertainty of models *through the lens* of annotators with different sociodemographics could help to identify social biases against vulnerable groups.

To this end, we exploit the conformal prediction framework to assess the uncertainty and reliability of model predictions. Unlike conventional approaches that prioritize accuracy, conformal prediction provides a metric for evaluating the alignment between model outputs and the confidence required for fair decision-making. By identifying disparities, this framework offers a structured approach to understanding biases, ultimately fostering fairness and inclusivity in model design and evaluation.

Specifically, we formulate two research ques-

[1]Code and data of our experiments are on https://github.com/aequa-tech/conformal-prediction-bias

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18061–18076
November 4-9, 2025

tions.

RQ 1. Is a models' uncertainty in automatic content moderation a predictor of biases against vulnerable groups to discrimination?

RQ 2. Can the fairness of models be assessed using user representations based on uncertainty?

Addressing these questions, we show that our approach brings out general patterns of hidden discrimination against non-white people but also shows that some street-level algorithms are expected to be fairer than others.

The main contributions of this research are the following: *i.* We introduce an unsupervised approach that leverages uncertainty to assess the fairness of models' predictions; *ii.* We provide a benchmarking analysis of 11 NLP systems that exhibit different levels of alignment with annotations provided by annotators belonging to vulnerable groups to discrimination; *iii.* We demonstrate that representing users through the uncertainty of model predictions is effective to observe the tendency of models to align with specific socio-demographic groups.

2 Related Work

Unlike structured domains where uncertainty estimation is often a standard consideration, NLP research has traditionally focused on maximizing accuracy-based evaluation metrics, such as F_1 score and log-likelihood, underestimating the computation of model uncertainty. Uncertainty measures reached, only recently, visibility in the NLP community (Xiao and Wang, 2019; Vázquez et al., 2024). NLP tasks, indeed, involve inherently ambiguous data, where human label variation introduces significant variability into annotated datasets. And, in such cases, conventional evaluation metrics fail to capture the full extent of model uncertainty, potentially leading to overconfident and unreliable predictions.

Bias detection. Scientific literature has revealed important biases coming from data created and annotated by specific segments of the population, leading to the creation of non-neutral models (Santurkar et al., 2023) and to the reinforcement of social stereotypes (Caliskan et al., 2017). Various techniques have been proposed to reveal biases and models' viewpoints: evaluation of contextualized word embeddings (Basta et al., 2019; Ethayarajh et al., 2019), specific evaluation frameworks (Barikeri et al., 2021; Felkner et al., 2023), questionnaires (Scherrer et al., 2023; Wright et al., 2024), transformer-based recognizers (Benkler et al., 2023), special prompts (Cao et al., 2023; Tao et al., 2024), and interaction with users (Shen et al., 2024; Kirk et al., 2024). Moreover, recent theoretical frameworks (Uma et al., 2021; Frenda et al., 2025) underline the need to take into account various *perspectives* about linguistic and pragmatic phenomena. Especially the detection of subjective phenomena (i.e., toxic language) proved to be affected by different perceptions that reflect annotators' backgrounds, beliefs, values, and identities (Sap et al., 2022; Fleisig et al., 2023). Therefore, a content moderation system should be representative of these different opinions, especially if these opinions come from segments of the population that are actually targets of attacks online (Kalluri, 2020). Focusing on toxic language detection, datasets like SBIC and CREhate with multiple annotations and information about annotators have been proposed and proved to be useful for investigating biases (as we have done in this work), building inclusive (Casola et al., 2023) and personalized models (Kocoń et al., 2021), and providing informative explanations about models' decisions (Mastromattei et al., 2022).

Confidence and multiple annotations. The most common method to estimate confidence in models is the logit-based method that assesses their uncertainty using token-level probabilities employable to LLMs (Geng et al., 2024) and other models (Wu and Klabjan, 2021). In Frenda et al. (2023), softmax-based measure of uncertainty has been employed to analyze the level of confidence of models trained on the annotations of specific segments of the population compared with a model trained on majority voting decision, showing that the formers tend to make a decision with less uncertainty than the standard model. Similar results are reported by Anand et al. (2024), where the use of Multi-Ground Truths models, trained on instance-annotator label pairs, improved confidence for samples characterized by substantial annotation disagreements. In this last work, the confidence is computed as the mean class probability for each data's gold label across the epochs.

Conformal Prediction in NLP. Differently from previous works, we rely on conformal prediction (Angelopoulos and Bates, 2023). This framework offers a systematic way to account for model

uncertainty, providing confidence measures that can inform decision-making, improve model interpretability, and mitigate biases in automated language processing systems. One of its key strengths is its ability to maintain robustness under distribution shifts, a property that has led to its widespread adoption in fields such as time series analysis (Shafer and Vovk, 2008; Papadopoulos, 2008). In many real-world applications, models are trained on datasets that may not fully capture the variability and complexity of the data they encounter in deployment. This discrepancy between training and deployment distributions—commonly referred to as distribution shift—can severely impact model performance and reliability. Conformal prediction helps mitigate this issue by providing statistically rigorous uncertainty estimates that remain valid even when the underlying data distribution changes (Vovk et al., 2005). This adaptability makes it particularly useful in settings where data evolves over time or where collecting perfectly representative training samples is infeasible, like in our case. Following its success in domains like time series forecasting, conformal prediction has been employed recently in NLP (Campos et al., 2024). Villate-Castillo et al. (2025), for instance, used conformal prediction in a framework of content moderation based on model uncertainty in predicting toxicity and disagreement among annotators. Differently from our work, the authors do not focus on highlighting demographic-based biases in models to detect hate speech.

3 Methodological Framework

Our methodology relies on conformal prediction, a statistical framework used to quantify the reliability of model predictions by assessing conformity, or how well individual predictions align with a set of labels (Angelopoulos and Bates, 2021). We leverage this theoretical framework to design two metrics for the analysis of bias in pre-trained models with the specific aim to measure their uncertainty against four socio-demographic groups based on the intersection of gender and ethnicity: white men, white women, non-white men, and non-white women.

3.1 Uncertainty Divergence

We utilized the *Brier Score* as a core component to implement a conformal prediction framework to compare the average uncertainty of a model against a given annotator and the gold standard label obtained through majority vote.

For a single annotated text, and a set of possible labels, $\mathcal{Y}$, the *Brier Score* $b(t_k)$ for text t_k can be written as

$$b(t_k, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \left(o_y(t_k) - p_M(y \mid t_k)\right)^2 \quad (1)$$

where:

- $o_y(t_k)$ is the binary indicator (1 if the true label is y, else 0).
- $p_M(y \mid t_k)$ is the model-predicted probability for label y.

The *Brier Score* is directly used as a single conformity score to quantify the alignment of model predictions with observed outcomes. A lower score indicates better conformity, reflecting predictions that are less uncertain and better calibrated.

Conformity delta. Since annotations often reflect the individual perspectives of annotators, beyond aggregated labels, we quantified prediction uncertainty by introducing the concept of the *Conformity Delta* (Δ). This measures the variability in the model's confidence when predictions are compared across individual and aggregated labels, providing deeper insights into uncertainty and reliability.

Let $a \in \mathcal{A} = \{a_1, .., a_m\}$ be a single annotator, $\mathcal{T} = \{t_1, \ldots, t_n\}$ denote the set of annotated texts, and M represent an automatic annotation model that outputs a probability distribution over labels $\mathcal{Y}$.

For a text $t_k \in \mathcal{T}$, let the label provided by the annotator a_i be $y_a \in \mathcal{Y}$, and given the ground-truth label y_A for the text t_k, obtained as the majority score among a specific subset of annotators $A \subset \mathcal{A}$ (e.g., the annotators belonging to a particular demographic group), the uncertainty $\delta(t_k)$ for the text t_k and the annotator a_i is defined as

$$\delta_{a_i}(t_k) = b(t_k, y_{a_i}) - b(t_k, y_A), \quad (2)$$

This $\delta_{a_i}(t_k)$ measures the variability in model confidence when predictions are evaluated against individual versus aggregated labels. A high value indicates significant disagreement or variability, often highlighting areas where annotators may have diverging perspectives or where model predictions fail to achieve consistency across groups.

Let Δ_A and Δ be defined as.

$$\Delta_A = \{\delta_{a_i}(t_k)\} \quad a_i \in A, \quad k = 1, \ldots, n \quad (3)$$

Finally, we can also consider the fully disaggregated Δ, being the set of all disaggregated labels:

$$\Delta_{\mathcal{A}} = \{\delta_{a_i}(t_k)\} \quad a_i \in \mathcal{A}, \quad k = 1, \ldots, n \quad (4)$$

Uncertainty divergence. The combination of Brier Score and Conformity Delta enables a nuanced assessment of model performance. While the Brier Score captures the overall prediction quality, the Conformity Delta highlights cases where individual annotators diverge significantly from the consensus label. The specific delta sets Δ_A and $\Delta_{\mathcal{A}}$ offer complementary advantages: Δ_A allows for targeted analysis of specific demographic groups to identify group-specific biases, while $\Delta_{\mathcal{A}}$ provides a comprehensive view across all annotators to capture the full spectrum of individual variations.

This divergence may indicate areas where the model struggles to generalize or where ground-truth labels are inherently ambiguous. By identifying such discrepancies, this approach enhances interpretability and guides iterative model refinement. This is particularly important in tasks like abusive content moderation, where decisions can amplify societal biases if models are not carefully evaluated. Discrepancies in annotator labels, influenced by cultural or personal factors, can lead to biased training data. By systematically quantifying uncertainty through the Conformity Delta, we can identify areas of potential bias and ensure that AI systems operate transparently and equitably.

In order to compute a potential correlation between the four socio-demographic groups of annotators and their average conformity deltas, we introduce the *Uncertainty Divergence*. For each group, we convert the obtained conformity deltas in Δ_A in a distribution with three categories: Conformity $\delta_{a_i}(t_k) < 0$, Conformity $\delta_{a_i}(t_k) = 0$, Conformity $\delta_{a_i}(t_k) > 0$. We compute the Kullback-Leibler divergence (van Erven and Harremos, 2014), which is defined as:

$$D_{KL}(P\|Q) = \sum_{i \in \{<0,0,>0\}} P(i) \log \frac{P(i)}{Q(i)} \quad (5)$$

Where P is the distribution of conformity Δ assigned to all the disaggregated labels in the corpus and Q the group-based one. For each model, we report the general uncertainty, under the label "total", defined as the average of the single uncertainties across the whole dataset. This enables measuring the general uncertainty of benchmarked models and specific uncertainties against socio-demographic groups.

3.2 Demographic Divergence.

To computationally represent annotators, we leveraged the concept of uncertainty derived from the computation of Conformity Delta (Equation 3). The uncertainty interval falls within the range $[-1, 1]$. Both ends of the scale indicate maximal disagreement, where the model strongly favors a label different from the one chosen by the annotator, albeit in opposite directions. A value of 0 indicates perfect agreement, where the model aligns with the annotator's label.

Each annotator $a \in \mathcal{A}$ is finally represented by a 40-dimensional vector $\mathbf{v}_a \in \mathbb{R}^{40}$, where each element $\mathbf{v}_a[j]$ corresponds to the frequency of uncertainty values δ for the texts annotated by a that fall within the j-th bin Bin_j. The number of bins was selected based on empirical experimentation: we varied the discretization granularity from 10 to 100 bins and observed the effect on clustering with KMeans. While the inertia naturally increases with higher-dimensional vectors due to geometric dispersion, the incremental gain in resolution starts to plateau around 40 bins, indicating that further increasing the number of bins does not substantially improve the discriminative power of the annotator representation.

Given that, we can define the value of the j-th element of the vector $\mathbf{v}_a$ for the model M as

$$\mathbf{v^M}_{a_i}[j] = \frac{1}{|\mathcal{T}_{a_i}|} \sum_{t_k \in \mathcal{T}_{a_i}} \mathbb{I}(\Delta_{a_i}(t_k) \in \text{Bin}_j), \quad (6)$$

where $\mathcal{T}_{a_i} \subseteq \mathcal{T}$ is the set of texts annotated by a_i, $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition is true and 0 otherwise, and Bin_j in one of the 40 equally sized bins defined as:

$$\text{Bin}_j = \left[-1 + \tfrac{(j-1)\cdot 2}{40}, -1 + \tfrac{j\cdot 2}{40}\right), \quad j = 1, \ldots, 40. \quad (7)$$

The resulting vector $\mathbf{v^M}_a$ effectively characterizes the annotator's judgment patterns relative to the model: high values in bins near 0 would indicate an annotator who frequently agrees with the model, high values in bins near -1 or 1 would indicate an annotator who often strongly disagrees

with the model, and overall the specific distribution across all 40 bins creates a unique "uncertainty fingerprint" for each annotator.

This representation allows for subsequent clustering of annotators based on similarities in their uncertainty profiles, providing insights into distinct annotation behaviors and potential subgroups within the annotator population.

Demographic Divergence. For a given model M, we assess how much the demographic class distributions vary across the four clusters. To quantify this, we compute the *Jensen-Shannon Divergence (JSD)* (Menéndez et al., 1997) (weighted on the cluster size) across the distributions of demographic classes in the clusters, treating the four cluster distributions as components of a mixture model. Let $\pi_1, \ldots, \pi_4$ be the weights of the clusters (where $||c_i||$ is the number of annotators inside the cluster i and $\pi_i = ||c_i|| / \sum_{j=1}^{4} ||c_i||$), and the demographic probability distributions inside the clusters are $P_1, \ldots, P_4$, for each language model, M, we define:

$$\text{JSD}_{\pi_1,\ldots,\pi_4}(P_1, \ldots, P_4) = \sum_{i=1}^{4} \pi_i D(P_i \parallel M) \quad (8)$$

If the clustering, solely based on annotators' uncertainty, does not show significant differences in demographic distributions across clusters, the model can be considered fair, as the uncertainty is not influenced by annotators' demographic characteristics.

4 Experimental Setup

In our experimental setting, we provide two studies that rely on the conformal prediction framework to assess the presence of bias in pre-trained models against four socio-demographic groups based on the intersection between gender and ethnicity: white men, white women, non-white men, and non-white women. The first study explores if the adoption of uncertainty can be a predictor of biases against vulnerable groups (RQ1) and leverages the *Uncertainty Divergence* (Section 3.1). The second study assesses models' fairness in user representation (RQ2) through the *Demographic Divergence* metric (Section 3.2). We analyze the uncertainty of 8 fine-tuned LMs and 3 LLMs in the classification of hate speech. We adopt as a benchmark two disaggregated corpora annotated for hate speech, which include information about annotators, such as gender and ethnicity. Therefore, we are able to measure the uncertainty of each model against specific communities of people.[2]

Pre-trained Models. In order to account for different generations of models, we considered for our benchmark study a set of transformer-based language models, including both fine-tuned LMs and prompted LLMs in a zero-shot setting. For fine-tuned LMs, we based our selection on NLP community adoption metrics: we included all models with at least $1,000$ downloads on HuggingFace[3] during November 2024. This criterion yielded 8 language models for our experiments, representing a broad spectrum of training methodologies. As a result, we identified 8 LMs for our experiment that have been trained with a wide range of approaches:

- IMSyPP (Kralj Novak et al., 2022). A BERT-based model (Devlin et al., 2019) trained on a multilingual corpus of hate speech messages gathered from Youtube and Twitter with disaggregated annotations.
- HateBert (Caselli et al., 2021). A retrained version of BERT based on RAL-E: a dataset of posts from banned SubReddits.
- Dynabench (Vidgen et al., 2021). A model trained on a dynamically annotated dataset in which messages have been annotated through a multi-step process.
- Twitter-Roberta-Base (Antypas and Camacho-Collados, 2023). A BERT-based model trained on a composition of 13 corpora annotated for hate speech, misogyny, and other correlated phenomena.
- Refugees[4]. A model developed as a collaboration between UNHCR, the UN Refugee Agency, and Copenhagen Business School.
- DistilRoberta (badmatr11x, 2023a). A fine-tuned version of RoBERTa on the badmatr11x dataset (badmatr11x, 2023b)
- Pysentimiento (Perez et al., 2023). Trained on the HatEval dataset (Garibo i Orts, 2019), the model is part of a multilingual toolkit developed for the detection of hate speech, sentiment analysis, emotion, and irony.

[2] All the experiments have been run on a RTX 3070 TI with the Hugging Face library *transformers*. We adopt the default setup of each model as it is available on Hugging Face.

[3] https://huggingface.co/

[4] henrystoll/hatespeech-refugees

- MuRIL (Das et al., 2022). This is a fine-tuned version of MuRIL on English abusive speech dataset.

Furthermore, we selected 3 open-source LLMs to replicate the experiment in a zero-shot setting.

- Mistral (Jiang et al., 2023) is one of the first European LLMs developed by a start-up led by former scientists of Facebook-AI.
- Olmo (Groeneveld et al., 2024) is the LLM of AllenAI and is characterized by the careful implementation of toxicity filtering strategies from the pretraining corpora.
- Bloom (Scao et al., 2022) is the outcome of a series of workshops that involved hundreds of NLP scientists.

Corpora. For our experiment, we chose two existing disaggregated corpora annotated for hate speech detection: the Social Bias Inference Corpus (SBIC) and CREHate. The rationale for choosing these resources is twofold: *i.* they represent two different generations of perspectivist datasets; *ii.* they significantly vary in their size and average number of annotations per message.

`SBIC` (Sap et al., 2020) is the first disaggregated corpus for hate speech detection that includes information about annotators' gender and ethnicity. The dataset consists of $44,671$ messages collected from multiple sources of previously annotated corpora for the same phenomenon. SBIC is composed of $146,254$ annotations with an average of 3.2 annotations per message. The number of individual labels diverging from the gold standard label represents the 4.9%

`CREHate` (Lee et al., 2024) is the latest disaggregated corpus for hate speech detection. CREHate is composed of $1,580$ messages from existing hate speech corpora that were re-annotated. The dataset includes $42,546$ annotations with an average of 26.9 annotations per message. The number of individual labels diverging from the gold standard label represents the 9.7%

Further description of both corpora and annotators is provided in Appendix A.

5 STUDY 1: Models Uncertainty towards Socio-Demographic Groups

In this study we adopt Uncertainty Divergence (Section 3.1) to compare models' performance with their uncertainty against texts labeled by annotators belonging to four socio-demographic groups: white men ($w.m.$), white women ($w.f.$), non-white men ($\neg w.m.$), and non-white women ($\neg w.f.$).

We compute the F_1 score obtained with each model against the total list of disaggregated labels and the F_1 score against labels of a specific group. This enables ranking the general performance of each model and observing differences between groups. Table 1 shows results of this analysis. As can be observed, the Refugees model obtains the best F_1 score on both SBIC (0.57) and CREHate datasets (0.55, shared with Olmo-7B). The analysis broken down by groups shows a distinction on the gender axis. In 15 cases out of 22, all models are better at predicting labels annotated by women. In SBIC, models perform a higher F_1 annotation of non-white women; in CREHate white women. A second pattern is about the performance of LLMs. In both cases, a pattern based on race emerges. Their predictions are better on non-white people in SBIC, while the opposite is observable for CREHate.

The results in Table 2 show that LLM predictions are more prone to uncertainty: 2 out 3 obtain the highest average conformity Δ. Observing the divergence of each group, it is possible to identify a systematic lower uncertainty in the classification of men's labels: white in SBIC, non-white in CREHate.

The study shows that Uncertainty Divergence might be a reliable metric to investigate social biases emerging in classification. The metric does not correlate with the performance of models **[RQ1]**. Computing the T-Test (Kim, 2015) between F_1 scores and conformity Δ's show that these two scores are not correlated both in SBIC ($p = 0.14$) and CREHate ($p = 0.11$). This suggests that models with a higher performance but a lower conformity might fail in understanding unseen messages outside corpora distribution. In this sense, the higher conformity Δ assigned to annotations provided by non-white people can be interpreted as **a predictor of a potential systematic misalignment between street-level algorithms decisions and women perception of hate speech**. Therefore, conformity might be not only used as a metric of fairness but as a guiding principle for selecting for content moderation models that are able to *see through the lens* of vulnerable minorities.

	F1 score									
	SBIC					CREHate				
model	**total**	$w.m.$	$w.f.$	$\neg w.m.$	$\neg w.f.$	**total**	$w.m.$	$w.f.$	$\neg w.m.$	$\neg w.f.$
IMSyPP	0.41	+58e-3	-17e-3	-4e-2	+1e-2	0.33	-16e-3	+1e-3	-7e-4	+6e-3
HateBert	0.51	-31e-4	-23e-4	55e-4	55e-4	0.49	-2e-3	+1e-3	+1e-3	+3e-4
Dynabench	0.29	-5e-3	+5e-3	-26e-3	+35e-3	0.34	-4e-4	+13e-3	+8e-3	+5e-3
Twitter-Roberta-Base	0.31	+1e-3	+8e-3	-19e-3	+27e-3	0.37	-3e-3	-1e-2	+4e-3	+3e-3
Refugees	**0.57**	-7e-3	+6e-3	+29e-3	-9e-3	**0.55**	+4e-3	+2e-3	-38e-4	-24e-4
DistilRoberta	0.44	-11e-3	+6e-3	-19e-3	+1e-2	0.44	-1e-2	+9e-3	+6e-4	-4e-4
Pysentimiento	0.35	8e-3	2e-3	-58e-4	+26e-3	0.33	-19e-3	+13e-3	+13e-3	+9e-3
MuRIL	0.36	10e-3	+12e-6	-55e-4	12e-3	0.31	-17e-3	+12e-3	+6e-4	+5e-4
Olmo-7B	0.48	-68e-4	+35e-5	+51e-3	-12e-3	0.55	12e-3	-44e-4	-74e-4	-21e-4
Bloom-7B	0.48	-30e-4	-14e-4	+19e-3	+16e-3	0.49	+22e-4	-65e-4	+6e-4	+11e-4
Mistral-7B	0.49	-23e-4	+31e-4	-21e-4	+53e-4	0.50	-67e-4	+75e-4	+24e-4	-40e-4

Table 1: Delta F_1 score for SBIC and CREHate obtained with each model against the total list of disaggregated labels and the F_1 score against labels of a specific group: white men ($w.m.$); white women ($w.f.$); non-white men ($\neg w.m.$), non-white women ($\neg w.f.$). Models that encode optimally the perspective of a group are highlighted in green. Models that predict worse on the group than on majority voting are highlighted in red.

	Uncertainty Divergence									
	SBIC					CREHate				
model	**total**	$w.m.$	$w.f.$	$\neg w.m.$	$\neg w.f.$	**total**	$w.m.$	$w.f.$	$\neg w.m.$	$\neg w.f.$
IMSyPP	-3e-4	55e-5	88e-5	10e-3	66e-4	-19e-4	19e-4	12e-4	29e-5	16e-4
HateBert	-12e-4	49e-6	67e-5	34e-4	45e-4	-5e-4	83e-5	68e-5	40e-5	16e-4
Dynabench	-7e-4	59e-6	76e-5	36e-4	45e-5	-9e-4	84e-5	88e-5	32e-5	18e-4
Twitter-Roberta-Base	-15e-4	15e-5	10e-5	45e-4	45e-4	13e-4	19e-5	13e-5	10e-3	66e-4
Refugees	13e-4	19e-5	13e-5	10e-3	66e-4	4e-4	11e-4	55e-5	39e-5	15e-4
DistilRoberta	-3e-4	43e-6	70e-5	35e-4	46e-4	-12e-4	94e-5	72e-5	29e-5	16e-4
Pysentimiento	14e-4	91e-5	20e-4	58e-4	47e-4	-23e-4	23e-4	12e-4	30e-5	16e-4
MuRIL	-71e-4	10e-4	20e-4	59e-4	49e-4	19e-5	23e-4	12e-4	31e-5	16e-4
Olmo-7B	16e-4	54e-5	18e-4	10e-3	30e-4	56e-5	13e-4	6e-4	4e-4	15e-4
Bloom-7B	16e-4	15e-5	10e-4	39e-4	42e-4	32e-4	14e-4	9e-4	3e-4	15e-4
Mistral-7B	-14e-4	84e-6	10e-4	39e-4	42e-4	10e-4	8e-4	6e-4	3e-4	15e-4

Table 2: The Uncertainty Divergence between the distribution of non-conformity scores assigned by models to all the individual annotations against the distributions of annotations grouped by the intersection of gender and ethnicity. The higher the score (red), the higher the divergence between the non-conformity of all the individual annotations and socio-demographic groups. The lower the score (green), the lower the divergence.

6 STUDY 2: User representation: Clustering annotators according to their uncertainty

This study leverages Demographic Divergence to represent users through models' uncertainty in text classification. The annotators were grouped into four distinct clusters for each model M as described in Section 3.2. The number of clusters was set to 4 to align with the examined socio-demographic groups ($w.m.$, $w.f.$, $\neg w.m.$, $\neg w.f.$). Clustering was performed on the annotator vectors $\mathbf{v^M}_a, \forall a \in A$, which represent the distribution of uncertainty for each annotator, using k-means. Figure 1 shows an example of results based on HateBert uncertainty scores over the CREHate dataset. Bar plots represent the demographic distribution of annotators in each cluster, and line plots represent the uncertainty related to each socio-demographic group. All the clusters are available in Appendix B.

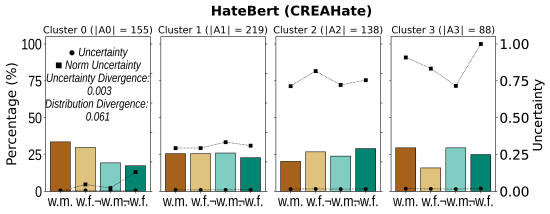


Figure 1: An example of annotator clusters based on the uncertainty of HateBert over the CREHate dataset.

At this stage, for each of the 11 models considered, we have two clusterings (one for the CREHate dataset and one for the SBIC dataset), each consisting of 4 clusters. Our goal is twofold: on the one hand, to quantify the distribution of demographic characteristics of the annotators within the clusters, and on the other, to evaluate Uncertainty

Divergence within individual clusters across the models.

For a model M and the four identified annotator clusters, we calculate the average uncertainty across demographic classes for each cluster by leveraging Uncertainty Divergence metrics. If the average uncertainty varies across the different clusters, it indicates that there are groups of annotators, for which the model M exhibits lower fairness compared to others.

Figure 2 illustrates the ranking of models based on Uncertainty Divergence and Demographic Divergence. Uncertainty Divergence and Demographic Divergence are specifically used to quantify the fairness of models **[RQ2]**. By examining the figure, we can see that the LLMs exhibit higher Uncertainty Divergence for both datasets. Therefore, despite achieving optimal results in terms of F_1, the higher uncertainty compared to the LMs indicates that there are groups of annotators for which LLMs exhibit lower fairness.

In contrast, Demographic Divergence helps us to understand whether this uncertainty is distributed equitably across the clusters, and hence, across the demographic classes. LLMs perform better in the ranking, particularly Mistral-7b. Consistent with Tables 1 and 2, this LLM maintains fairness across the considered dimensions (gender and ethnicity). On the other hand, Olmo-7b presents negative Demographic Divergence values, indicating higher uncertainty, which is not evenly distributed across the demographic classes.

We want to highlight the behavior of MuRIL, which exhibits the lowest uncertainty among all models for both datasets but presents the highest Demographic Divergence. In Appendix B, Figure P shows that clusters with female or non-white female annotators exhibit higher uncertainty, and the distributions between clusters are significantly different.

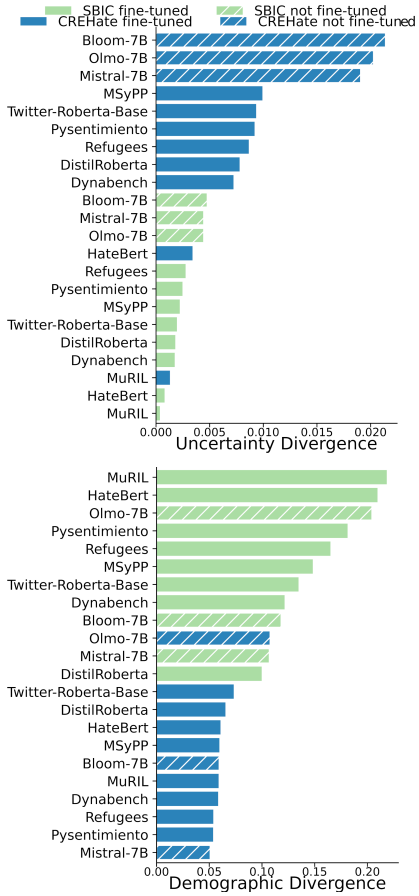


Figure 2: Ranking of the models based on Uncertainty Divergence and Demographic Divergence.

7 Discussion

Our research shows that uncertainty and the conformal prediction framework are **a powerful approach for the analysis of models' fairness in automatic content moderation**. Uncertainty is effective for the observation of the alignment of models' behavior with the sensibility of specific groups of annotators. Uncertainty is also a powerful way to understand how single annotators are represented and grouped, and to what extent models perpetuate social biases during this process.

From the benchmark analysis of pre-trained models, both general patterns and model-specific behaviors emerge. The vast majority of models show the lowest uncertainty in predicting contents annotated by men and the highest uncertainty for the prediction of contents annotated by non-white people. This implies that in the automatic content moderation settings, in which the model performance might suffer a high degradation due to temporal semantic shifts, the risk of misalignment between the model's predictions and human annotations is higher for non-white people. An explanation of such a pattern might be in the long-term impact of the pretraining process. Models trained on data that poorly represent non-white people learn perspectives of the world that are not easy to be removed. In this sense, the more nuanced results in terms of performance might be interpreted as an optimization over specific benchmark corpora, while **uncertainty could be the blueprint of pretraining biases**.

Despite the presence of common behaviors,

choosing one pre-trained model over another matters. The average uncertainty of these models significantly varies without correlating with performance (Section 5), showing that the two ways of measuring them (F_1 score and Conformity Δ) grasp different aspects of the same problem. In this sense **Mistral represents the best trade-off between performance and uncertainty**, despite the lack of information about which data has been used to train it (Jiang et al., 2023). However, it is not possible to draw conclusions from the comparison of fine-tuning and zero-shot approaches, since OLMO and Bloom exhibit a higher Demographic Divergence (Section 6) within their clusters: a predictor of bias against vulnerable groups to discrimination.

Corpora themselves appear to be a factor in stressing the uncertainty of models and represent a significant limitation for fairness studies. Model predictions on SBIC systematically suffer higher Demographic Divergence (Appendix B) than predictions on CREHate. The very different composition of annotators (Section 4) might play a role in this, as well as the different degrees of subjectivity emerging from raw annotations (4.9% *vs* 9.7%), and the average number of annotations per message (3.2 *vs* 26.9). Furthermore, the inter-annotator agreement between annotators belonging to different demographic classes is generally higher within the same class, while agreement between different classes tends to vary (Appendix C). Despite these differences, **the two corpora share the same limitation: binarism in the annotators' selection process**. Non-binary people were almost not involved in the annotation process, hindering most insightful analyses that go beyond the traditional intersection of race and gender. Resources developed for fairness should be more effective in representing marginalized and invisible groups of people.

8 Conclusion and Future Work

In this paper, we presented a novel approach to assess the fairness of models through their uncertainty. We introduced metrics for measuring the impact of uncertainty against socio-demographic groups. In particular, we leveraged our unsupervised approach based on conformal prediction to benchmark 11 street-level algorithms on SBIC and CREHate datasets: 8 LMs fine-tuned for hate speech detection and 3 LLMs instantiated through a prompt-based method. The results show that measuring models' uncertainty unfolds systematic and hidden biases against non-white people, which do not emerge from performance-based metrics, such as the F_1 score [**RQ1**].

Moreover, we generated vector representations of annotators based on uncertainty scores emerging from models' predictions and used them to cluster annotators. The socio-demographic composition of the resulting clusters significantly varies between models, which show different degrees of fairness against women and non-white people [**RQ2**].

Future work goes in two directions. We will test the impact of considering uncertainty during fine-tuning and active learning (e.g., through Reinforcement Learning approach) to reduce bias in model prediction. We will explore the transferability of our methodology on contiguous tasks to hate speech detection and to other perspectivist corpora.

Limitations

In this work, our approach has been tested on hate speech detection; however, to validate its generalizability, we will further employ it on the detection of other subjective phenomena (i.e., when a higher human label variation is a sign of diverse subjectivities). Additionally, we only choose a subset of models for our analysis. This might result in overlooking models that actually show different patterns in the representation of vulnerable groups than the ones emerging in our analysis. Finally, we focus in particular on dimensions of gender and ethnicity common to both datasets used as samples for proving our methodology. However, we are aware that a binary classification for gender and ethnicity is far from the real world and could raise discussion (Larson, 2017). Moreover, considering other identity axes, it is possible that hidden forms of discrimination could emerge. Nevertheless, our approach can be used with multiple categories across various dimensions.

Ethical Issues

Since this research relies on secondary data, there are no ethical issues related to the collection and annotation of texts. Research biases related to these previous studies may still have an impact on the representation of human annotators emerging from our results.

Acknowledgments

The work of S. Frenda is supported by the EPSRC project "Equally Safe Online" (EP/W025493/1).

References

Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Abhishek Anand, Negar Mokhberian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511.

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

badmatr11x. 2023a. Distilroberta-base-offensive-hateful-speech-text-multiclassification. https://huggingface.co/badmatr11x/distilroberta-base-offensive-hateful-speech-text-multiclassification.

badmatr11x. 2023b. Hate-offensive speech dataset. https://huggingface.co/datasets/badmatr11x/hate-offensive-speech.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *CoRR*, abs/2312.10075.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 32–42, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Virginia Dignum. 2023. Responsible artificial intelligence — from principles to practice: A keynote at thewebconf 2022. *SIGIR Forum*, 56(01).

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in

large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347.

Òscar Garibo i Orts. 2019. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. In *Annual Meeting of the Association for Computational Linguistics*.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(9), pages 13669–13672.

Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169.

Tae Kyun Kim. 2015. T test as a parametric statistic. *kja*, 68(6):540–546.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695, Cham. Springer International Publishing.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Michele Mastromattei, Valerio Basile, and Fabio Massimo Zanzotto. 2022. Change my mind: How syntax-based Hate Speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 117–125, Marseille, France. European Language Resources Association.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Harris Papadopoulos. 2008. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*, Rijeka. IntechOpen.

Juan Manuel Perez, Mariela Rajngewerc, Juan Carlos Giudici, Damián Ariel Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. pysentimiento: A python toolkit for opinion mining and social nlp tasks. *Research Square*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 30 others. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, and 5 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *Preprint*, arXiv:2406.09264.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Tim van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe, editors. 2024. *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*. Association for Computational Linguistics, St Julians, Malta.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz. 2025. A collaborative content moderation framework for toxicity detection based on multitask neural networks and conformal estimates of annotation disagreement. *Neurocomputing*, 647:130542.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer US, Boston, MA.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 324–335, New York, NY, USA. Association for Computing Machinery.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values

and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Huiyu Wu and Diego Klabjan. 2021. Logit-based Uncertainty Measure in Classification . In *2021 IEEE International Conference on Big Data (Big Data)*, pages 948–956, Los Alamitos, CA, USA. IEEE Computer Society.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329.

A Corpora and annotators description

Both datasets include information on the annotators' gender (male, female, non-binary) and ethnicity (White, Hispanic, Asian, etc.). Figure 3 illustrates the distribution of the annotators' demographic classes. Non-binary individuals were excluded due to the low number of annotators in this gender category. For ethnicity, we grouped white and non-white annotators separately to achieve a more balanced distribution. Gender and ethnicity were then combined to form four distinct demographic classes: white man ($w.m$), white female ($w.f$), not-white male ($\neg w.m$), non-white female ($\neg w.f$) (Figure 3).

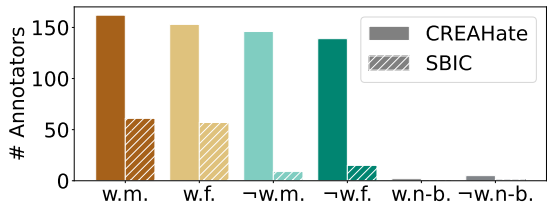


Figure 3: Annotators' demographics on CREHate and SBIC.

We excluded annotators who annotated fewer than 20 messages for two reasons. First, with too few messages, the uncertainty profile could be less reliable. Second, the threshold of 20 was chosen because the annotator with the fewest annotations in the SBIC dataset had annotated 24 messages. This approach ensured that, although the annotation distributions differed between datasets, they were made more comparable (Figure 4).

B Uncertainty in LLMs and LMs

This appendix provides a fine-grained analysis of the findings discussed in Section 6. For each model,

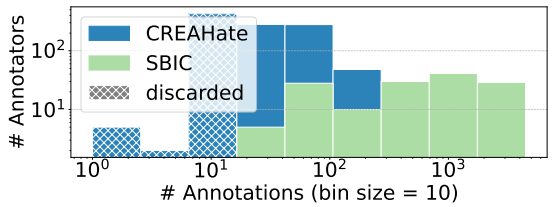


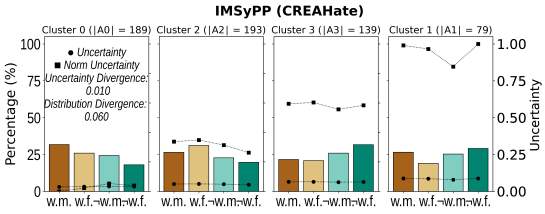
Figure 4: Distribution of the number of annotations per annotator.

we visualize the distribution of annotators' demographic classes alongside their corresponding uncertainty levels. Each row corresponds to a single model, with results from CREAHate on the left and SBIC on the right. Subplots $[a–o]$ display the results for LMs, while $[s–v]$ corresponds to LLMs. Notably, the SBIC dataset often results in very small annotator clusters. However, this does not affect the Uncertainty Divergence and Demographic Divergence metrics, as these are weighted by the number of annotators in each cluster. A key observation is that uncertainty—particularly when considering its normalized values across the four clusters—tends to be higher in clusters where women or non-white annotators are more prevalent (Figure 4).

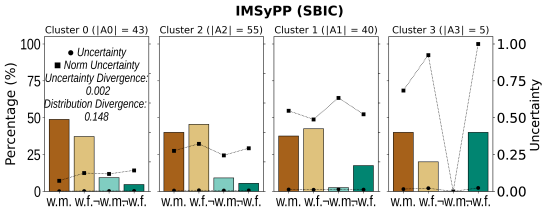
C Cohen's Kappa between annotators pairs

The following heatmaps show the pairwise Cohen's Kappa agreement between annotator classes for the two datasets analyzed in this study (SBIC and CREAHAte). Each matrix displays agreement values between the four demographic groups.

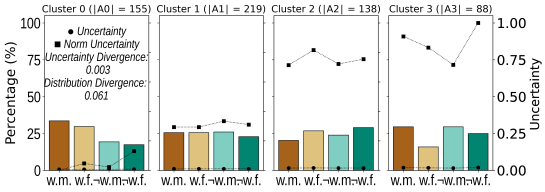
As expected, higher agreement is observed along the diagonal, indicating that annotators within the same class tend to be more consistent in their annotations. Off-diagonal values represent inter-class agreement, which is generally lower, highlighting differences in annotation behavior across demographic groups. These visualizations provide a detailed view of intra- and inter-class consistency and help contextualize the results reported in the main text.


IMSyPP (CREAHate)
Cluster 0 (|A0| = 189)
Cluster 2 (|A2| = 193)
Cluster 3 (|A3| = 139)
Cluster 1 (|A1| = 79)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.010
Distribution Divergence: 0.060
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

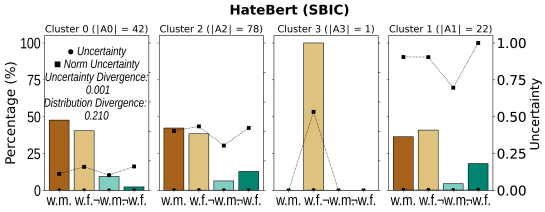
(a)


IMSyPP (SBIC)
Cluster 0 (|A0| = 43)
Cluster 2 (|A2| = 55)
Cluster 1 (|A1| = 40)
Cluster 3 (|A3| = 5)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.002
Distribution Divergence: 0.148
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

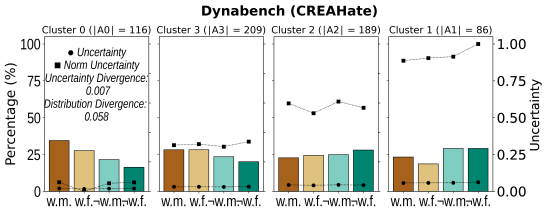
(b)


HateBert (CREAHate)
Cluster 0 (|A0| = 155)
Cluster 1 (|A1| = 219)
Cluster 2 (|A2| = 138)
Cluster 3 (|A3| = 88)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.003
Distribution Divergence: 0.061
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

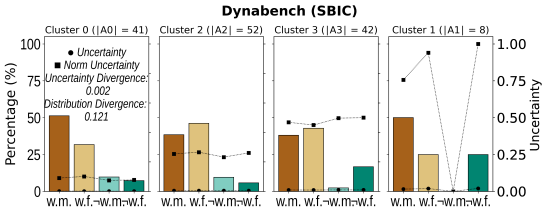
(c)


HateBert (SBIC)
Cluster 0 (|A0| = 42)
Cluster 2 (|A2| = 78)
Cluster 3 (|A3| = 1)
Cluster 1 (|A1| = 22)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.001
Distribution Divergence: 0.210
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

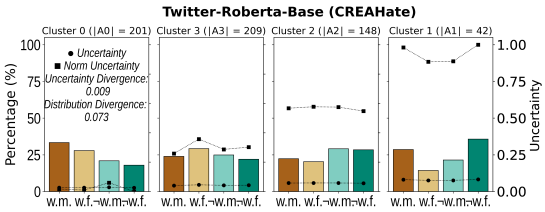
(d)


Dynabench (CREAHate)
Cluster 0 (|A0| = 116)
Cluster 3 (|A3| = 209)
Cluster 2 (|A2| = 189)
Cluster 1 (|A1| = 86)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.007
Distribution Divergence: 0.058
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

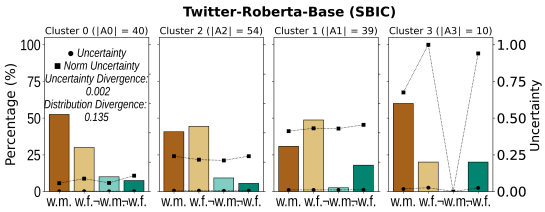
(e)


Dynabench (SBIC)
Cluster 0 (|A0| = 41)
Cluster 2 (|A2| = 52)
Cluster 3 (|A3| = 42)
Cluster 1 (|A1| = 8)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.002
Distribution Divergence: 0.121
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

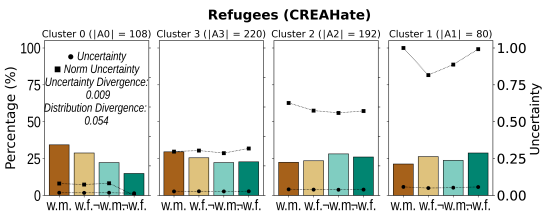
(f)


Twitter-Roberta-Base (CREAHate)
Cluster 0 (|A0| = 201)
Cluster 3 (|A3| = 209)
Cluster 2 (|A2| = 148)
Cluster 1 (|A1| = 42)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.009
Distribution Divergence: 0.073
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

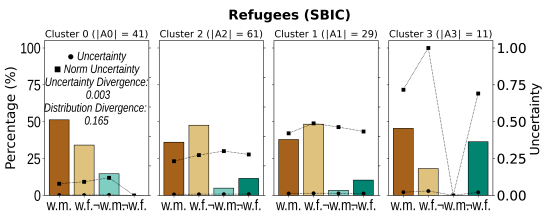
(g)


Twitter-Roberta-Base (SBIC)
Cluster 0 (|A0| = 40)
Cluster 2 (|A2| = 54)
Cluster 1 (|A1| = 39)
Cluster 3 (|A3| = 10)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.002
Distribution Divergence: 0.135
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

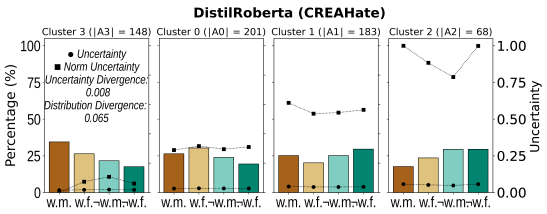
(h)


Refugees (CREAHate)
Cluster 0 (|A0| = 108)
Cluster 3 (|A3| = 220)
Cluster 2 (|A2| = 192)
Cluster 1 (|A1| = 80)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.009
Distribution Divergence: 0.054
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

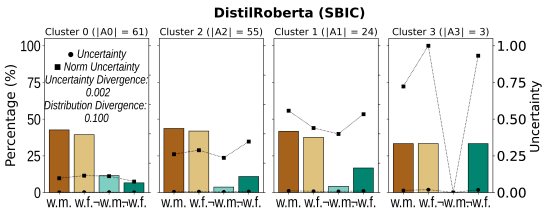
(i)


Refugees (SBIC)
Cluster 0 (|A0| = 41)
Cluster 2 (|A2| = 61)
Cluster 1 (|A1| = 29)
Cluster 3 (|A3| = 11)
Uncertainty
Norm Uncertainty
Uncertainty Divergence: 0.003
Distribution Divergence: 0.165
Percentage (%)
Uncertainty
w.m. w.f.¬w.m¬w.f.

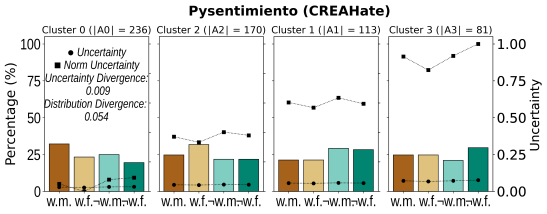
(j)



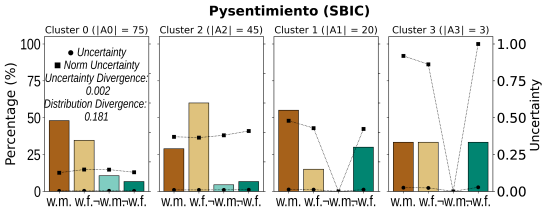
(k)



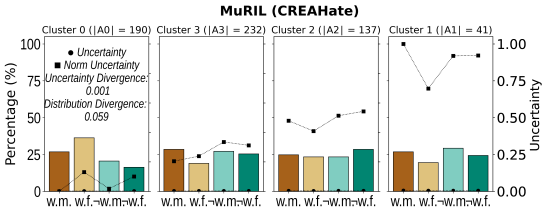
(l)



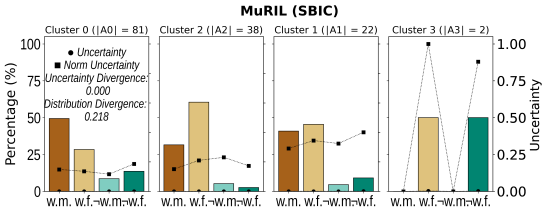
(m)



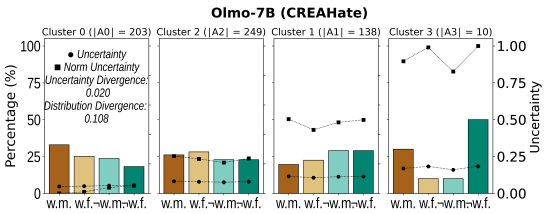
(n)



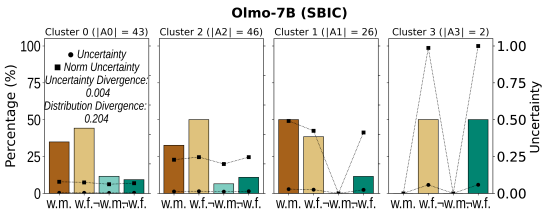
(o)



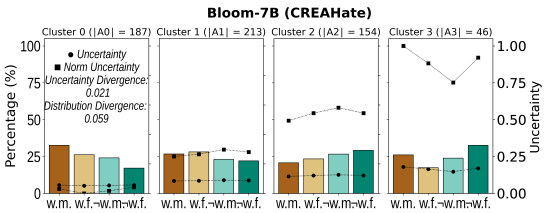
(p)



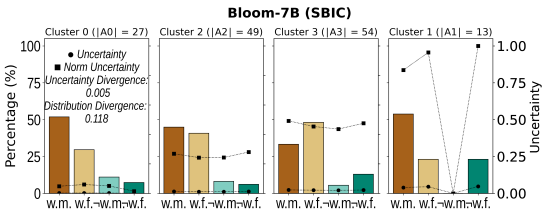
(q)



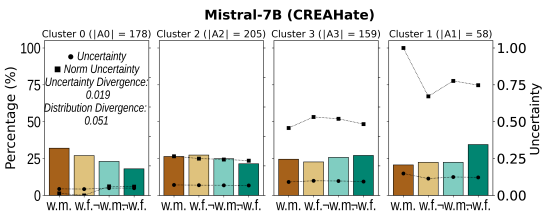
(r)



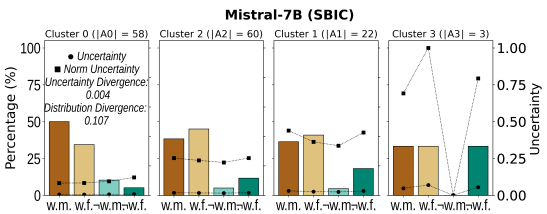
(s)



(t)



(u)



(v)

Figure 4: There are 22 plots, corresponding to the evaluation of 11 models on two distinct datasets, CREAHate and SBIC. Each plot consists of four subplots, where each subplot represents a cluster and illustrates the demographic distribution of the annotators within that cluster. Additionally, the subplots display the level of uncertainty, including its normalized value across the four clusters.

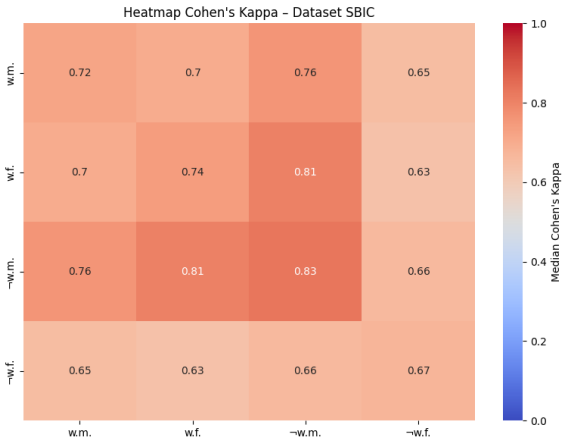


Figure 5: Pairwise Cohen's Kappa agreement between annotator classes per SBIC.

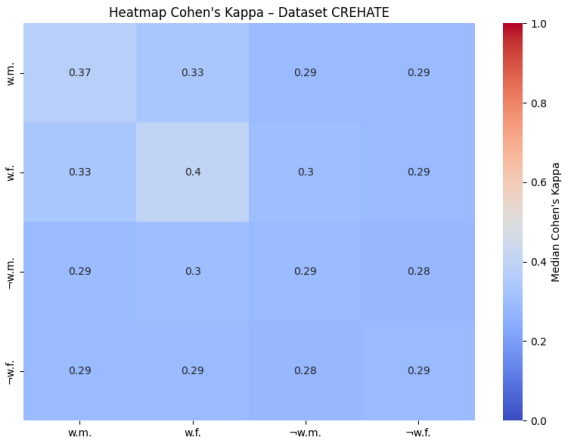


Figure 6: Pairwise Cohen's Kappa agreement between annotator classes per CREAHate.