Testing LLMs' Sensitivity to Sociodemographics in Offensive Speech Detection

Lia Draetta[1,*,†], Soda Marem Lo[1,*,†], Samuele D'Avenia[1,*,†], Valerio Basile[1] and Rossana Damiano[1]

[1]*Computer Science Department, University of Turin, Turin, Italy*

Abstract

Recent research in text classification increasingly leverages generative Large Language Models (LLMs) to address a wide range of tasks, including those involving highly subjective linguistic phenomena, such as hate speech and offensive language detection, areas closely tied to semantics and pragmatics. A growing body of works in the NLP community is examining how annotators' backgrounds influence labeling decisions, while also studying model biases and alignment with different social groups. A frequently used technique with generative models is sociodemographic prompting: where LLMs are asked to impersonate individuals based on their known demographic traits. In this work, we further explore this technique and its limitations on a disaggregated dataset of offensive speech detection. We selected five models of 7 to 8 billion parameters, and prompted them to classify the sentences, providing all possible combinations of the available sociodemographic traits (gender, race and political leaning). Additionally, we prompted the models to provide brief explanations of their choices to investigate their motivations. Through both a consistent quantitative and qualitative analysis, we observed limitations in their ability to exploit demographic information. Results underscore the need for in-depth analysis going beyond performance metrics when this technique is adopted.

Keywords

Data perspectivism, Sociodemographic prompting, Offensive speech detection,

1. Introduction

The capabilities of Large Language Models (LLMs) are being increasingly assessed across a variety of tasks, reaching a point where they are now frequently used for annotation purposes as well. Some studies have shown promising results [1, 2], especially given the challenges of building large annotated corpora and the high costs associated with manual annotation. On the other hand, the use of LLMs for tasks involving highly subjective judgments [3], such as detecting hate speech and abusive language, has raised concerns about their ability to align with diverse perspectives and produce annotations that actually mirror human label choices [4, 5]. Moreover, Santy et al. [6] proposed a framework for quantifying the positionality of datasets and models by computing the correlation with demographic groups, uncovering a strong alignment with the WEIRD (Western, Educated, Industrialized, Rich, Democratic) population.

Inspired by the *perspectivist* approach — which treats disagreement as a valuable source of information rather than noise [7] — the NLP community has become increasingly attentive to the subjectivity inherent in highly opinionated and context-dependent tasks, such as hate speech detection [8]. To explore how people's views shift and what factors might be useful predictors of such diversity in annotators' labeling, numerous studies have examined the influence of raters' sociodemographic characteristics and cultural background [9]. While this approach often resulted in being effective with classification models [10, 11], when generative models are prompted with demographic information, results are mixed.

Identity-Aware AI workshop at 28th European Conference on Artificial Intelligence, October 25, 2025, Bologna, Italy

*Corresponding author.

†These authors contributed equally.

lia.draetta@unito.it (L. Draetta); sodamarem.lo@unito.it (S. M. Lo); samuele.davenia@unito.it (S. D'Avenia); valerio.basile@unito.it (V. Basile); rossana.damiano@unito.it (R. Damiano)

0009-0004-6479-5882 (L. Draetta); 0000-0002-5810-0093 (S. M. Lo); 0009-0000-9256-0707 (S. D'Avenia); 0000-0001-8110-6832 (V. Basile); 0000-0001-9866-2843 (R. Damiano)

Several recent studies have begun to investigate whether LLMs can adopt different perspectives during annotation, paving the way to a new prompting methodology known as sociodemographic prompting [12, 13]. By guiding the model to simulate the viewpoints of different social groups during the annotation process, researchers have attempted to uncover inherent biases while also highlighting both the potential and the limitations of using LLMs for data annotation and generation. There is no clear consensus about the effectiveness of this strategy, [14, 15] and it remains unclear how models adjust their outputs based on provided demographic information [16, 17].

Aiming to address this gap, in this study we present a systematic comparison of 5 models' outputs when prompted with different demographics and their intersections. We evaluated the LLMs on an off-the-shelf dataset focused on toxic speech detection, specifically for the presence of racist and offensive speech. The corpus also includes information about annotators' identities and beliefs, as well as textual characteristics [18]. The authors' results show a strong association between annotator identity and rating of toxicity, especially with respect to their political leaning: more conservative annotators were more likely to rate African American English (AAE) dialect as toxic. Building on these findings, we considered this dataset well-suited for our study, as it enables us to investigate both the effectiveness of sociodemographic prompting and how linguistic features of the text influence model predictions. To expand on previous studies, we prompt the models not only to produce an annotation for the input messages but also to generate an explanation for why that specific option was chosen. This allows us to gain more direct insights into the reasons why models assign particular labels, exploring whether the demographic information provided influences their interpretation.

In Hate Speech Detection, explanations are of crucial importance. In previous works, human annotators have been required to produce motivations why a certain text is offensive, and the model's capacity to produce them is also assessed [19]. Moreover, their usage for fine-tuning aids models in achieving a better performance [20]. A similar strategy is adopted in Humor Understanding, where generative models have been asked to explain jokes and articulate why specific caption-image combinations are perceived as funny, or to provide memes interpretations [21, 22]. In both fields, the quality of the explanations is assessed by comparing them to reference human ones. However, in this work, we are not interested in such comparison, since we leverage explanations as a first access to the reasons driving the models' decisions.

Specifically, we aim to answer the following research questions:

- **RQ1**: Does providing sociodemographic information to the LLM help in performing classification tasks linked to a subjective phenomenon of the language, such as offensive speech?
- **RQ2**: Which demographic features appear to be most influential in determining the assigned label for text classification?
- **RQ3**: Does sociodemographic information in the prompt alter the explanations generated by the model?

Our study highlights the challenges faced by selected models in addressing a complex phenomenon such as offensive speech in a social media context. On the one hand, we did not observe consistent performance improvements, regardless of whether sociodemographic traits were provided. On the other hand, both labeling behavior and generated explanations showed little to no variability across all settings. When considering labels and explanations together, it is evident how the models struggle to disambiguate cases where AAE, slurs or swearwords serve as markers of offensive content from those where they are used in a reclaimed, informal or ironic ways.

These findings raise important questions as to **whether sociodemographic prompting is an effective technique to model human annotation behavior in highly subjective and complex tasks**. Finally, they underscore the need for a careful and comprehensive error analysis when such a technique is adopted.

The paper is organized as follows: In Section 2 we present a review of the literature, in Section 3 we describe the dataset characteristics and how it has been used and filtered for this study. Section 4

outlines the experimental design and setup, while Section 5 presents and discusses the corresponding results. Finally, in Section 6, we provide both quantitative and qualitative analyses to assess the model's capacity to adapt its labeling behavior based on sociodemographic traits, alongside an error analysis and an examination of the explanations generated by the model.

The full code is available at the following link: https://github.com/liadraetta/intersectionality-llm.

2. Related works

2.1. LLMs alignments

The interest in how language models align with humans based on social and cultural backgrounds reflects long-standing concerns about bias and representation in NLP systems [23]. Given the strong generalization capabilities of language models, several studies have begun to highlight various encoded biases and their potential to reproduce and amplify such stereotypical associations [24].

More recently, a line of research has focused on investigating the prevailing worldviews that LLMs adopt when performing different tasks [25]. Looking at demographic alignment of LLMs during default tasks, a recent study assessed whether the predictions of LLMs align more closely with those of individuals from particular demographic groups when no explicit conditioning is applied [25]. The authors concluded that LLMs do not represent all segments of society equally, and that sociodemographic prompting systematically influences their outputs. Similarly, other studies have examined how different LLMs process various socio-economic dimensions, such as social class [26, 27] and religion [28], highlighting that these models often exhibit stratified and biased views. In this context, several works analyze LLMs' alignment with specific sociodemographic groups and show that model responses are biased towards responses by participants from Western countries [29, 12, 6].

2.2. Sociodemographic prompting

Studies in the field of data perspectivism have demonstrated that modeling annotators' views improves performance on subjective NLP tasks [7]. Recently, researchers have begun exploring the influence of sociodemographic prompting for classification tasks, aiming to understand how such information affects model behavior. Multiple studies have investigated how models adapt to provided demographic information, but univocal conclusions are still missing.

Schäfer et al. [25] claim that sociodemographic prompting, adopted on offensiveness and politeness ratings, influences results in a structured way, and their analyses show that LLMs exhibit variations in annotation based on demographic attributes. Beck et al. [17] found that sociodemographic prompting can improve zero-shot performance on subjective NLP tasks, however it does not consistently outperform standard prompting. Moreover, their results show that model variability was more strongly influenced by factors such as prompt formulation than by demographic attributes. Despite this, they find that sociodemographic prompting can be useful to identify ambiguous instances, thereby supporting annotation efforts rather than serving as an effective approach to data annotation.

When leveraging sociodemographic prompting to compare human biases with those exhibited by persona-based LLMs, models showed a limited capacity to modify their behavior and adapt to specific personas [30], reinforcing existing work that questions their ability to faithfully reproduce human behavior [12], especially sociodemographic behaviors [31]. On the other hand, strong alignment with White participants was observed when nine LLMs were evaluated on two subjective tasks (politeness and offensiveness) [32]. Moreover, the authors reported that sociodemographic prompting led to inconsistent improvements in the models' ability to process language from specific sub-populations. Finally, another study has also raised a general doubt about LLMs capacity to reflect diverse demographic traits [33], leading authors to caution against using LLM-based simulations for subjective tasks.

Prior work shows no consensus on the limits and potential of prompting with demographic traits for subjective tasks. Given these open issues, this study aims to shed light on how such traits influence model predictions and highlight these differences through explanations.

3. Dataset

The dataset used for conducting our experiment is the Annotators with Attitudes dataset, designed to explore the *who*, *why* and *what* of toxicity annotation [18]. The authors conducted two online studies, the former on a controlled set of 15 posts, and the latter on a larger set of posts, simulating a crowd-sourced dataset. For our research, we leveraged the second to make our results comparable with other studies in the field.

The corpus consists of 627 texts, annotated by 184 people from the US. The authors collected 3,463 annotations, released in a disaggregated fashion. Each post was annotated by a median of 6 people, balanced across politics and race: two white conservatives, two white liberals and two black participants. This distribution reflects the composition of the annotator pool, which was unbalanced with respect to the race trait.

Information about annotators' gender, politics, race and age was released to gain insights into *who* annotates. To explore the *why* of their rating, the authors collected annotators' beliefs in terms of seven attitude dimensions. Finally, each text was defined by three categories useful to investigate *what* is considered toxic, i.e. characteristics that tend to influence toxicity labeling: anti-black language, presence of African American English dialect (AAE), and vulgar language (e.g. swearwords, slurs).

The annotation process consisted of providing a score from 1 to 5 assessing how *racist* and how *offensive* the text was perceived to be, either personally ("to you") or generally ("to any"). Offensiveness scores were then averaged to produce a single rating. For our study, we focus on offensiveness by converting the averaged scores into a binary label. Scores below 3 were treated as Not Offensive, those above 3 as Offensive, while the 287 instances with a score exactly equal to 3 were removed. The impact of this filtering step was limited, as most annotations were concentrated at the extremes, with intermediate scores occurring far less frequently. The full distribution of scores is reported in Section A.

Coherently with the authors' analysis, we opted to work with three sociodemographic traits: gender (man, woman), race (black, white) and political leaning (conservative, liberal, neutral).[1] The distribution of annotators' races was unbalanced: 141 participants self-identified as White, 38 as Black, one person preferred not to disclose this information, and one each identified as Middle Eastern, Hispanic, Native, or Other. To address this imbalance, we excluded the 5 participants who did not self-identify as either White or Black, resulting in a dataset of 176 raters and 627 texts, each annotated with a median of 5. This yielded a total of 3,094 annotations, with 1,592 labeled as Offensive, and 1,502 as Not Offensive.

4. Methodology

This work investigates whether different generative LLMs can modulate their predictions based on sociodemographic information provided as input, and whether such prompting improves their performances (RQ1). Specifically, the model is asked to decide whether a text is offensive and justify its decision with an explanation.

As a first step, inspired by Balestrucci et al. [34], we tested two prompting strategies for our baseline (without providing demographic data as input). The first required the model to answer by assigning a binary label and then provide an explanation (denoted "Answer then Explain" *A-Ex*), while the second required the model to reason and produce an explanation first and then answer (denoted "Explain then Answer" *Ex-A*). Once the stronger approach was identified, we extended the experiments by introducing varying levels of sociodemographic information into the prompt, enabling us to investigate improvements in classification performance (RQ1), investigate the most influential sociodemographic traits (RQ2) and their usage in explanations generation (RQ3).

All experiments were conducted using a few-shot approach, with the model being provided with some examples with the output in the expected format.[2]

[1] In the original dataset, political leaning is represented on a scale from -1 to 1, where -1 indicates a left-leaning orientation, 1 a right-leaning one. Unlike the analysis in the referenced paper, which distinguishes only between conservative and liberal, we also assign a value of 0 to denote a neutral political stance.

[2] The selected texts were removed from the evaluation.

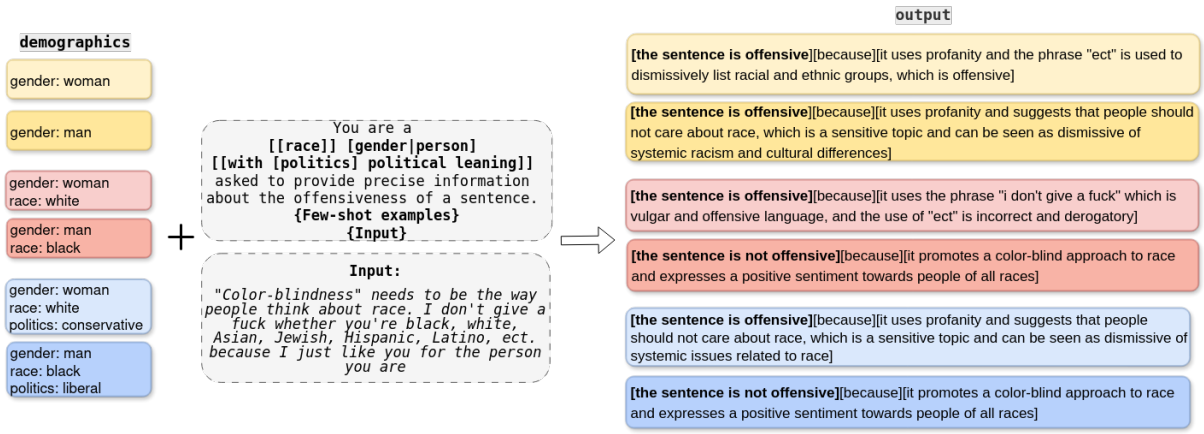

Figure 1: LLM annotations with sociodemographic traits.

We selected five decoder-only instruction-tuned models of comparable size (around 7 to 8 billion parameters): deepseek-llm-7b-base [35], Llama-3.1-8B-Instruct [36], Qwen2-7B-Instruct [37], Mistral-8B-Instruct-2410 [38], and gemma-7b-it [39]. The models were accessed via the Hugging Face API and run locally using a single NVIDIA A100 40GB with temperature 0.

4.1. Ex-A and A-Ex prompt construction

For the first set of experiments, we run all the models with the Ex-A and A-Ex approaches to identify the best one without any sociodemographic variables. When working with the A-Ex approach, the expected answer format is: "`[the sentence is offensive/is not offensive] [because] [explanation]`" while for the Ex-A approach it is: "`[reasoning explanation] [so] [the sentence is offensive/is not offensive]`".

The few-shot prompt was designed by including 4 input examples along with the output in the expected format. The dataset employed in this study is disaggregated, meaning that identical instances can have differing annotations. To avoid introducing bias by manually selecting the examples, we focused on sentences annotated by the maximum number of annotators (six people) with unanimous agreement among the annotators. Hence, we selected two sentences annotated as Offensive and two as Not Offensive. Additionally, two of the authors of the article, with a background in Linguistics, produced the explanations. The few-shot examples for both approaches are included in Appendix B.

4.2. Evaluation

Following the Perspectivist approach, we did not harmonize annotations into a gold standard. Instead, we preserved all individual annotations, treating each data point as a <text, annotator> pair. The LLMs were asked to assign a label to each pair, and their predictions were evaluated against the full disaggregated dataset described in Section 3. To measure their performance, we report standard classification metrics: Precision, Recall and F1 score.

In the following sections, when we refer to the "true labels", we mean the labels assigned by the individual annotators to the annotated texts.

4.3. Inclusion of sociodemographic information

Since the A-Ex approach yielded most of the top-performing models, including the best overall model (Section 5.1), the investigation on the influence of sociodemographic traits is conducted using this strategy.

In the second experiment, the annotators' sociodemographic profiles were included in the prompt, enabling an analysis of how identity framing may affect the model's predictions. The gender, race,

and political orientation variables were included in the prompt by asking the model to adopt a specific identity and respond from that perspective when performing the task. For example, if all three traits are specified (intersectional model), it takes the structure "*You are a [race] [gender] with [political view] political leaning asked to provide precise information about the offensiveness of a sentence*". If some variables are not specified, they are dropped from the persona description. In total, we designed 7 prompting conditions per model, corresponding to all the possible combinations of the demographic traits.

The full structure of the prompt, along with some of the generations obtained with varying amounts of sociodemographic context, is shown in Figure 1 with the full prompt in Appendix B.

5. Results

This section presents the model's performance on the classification task, beginning with a preliminary comparison of the A-Ex and Ex-A strategies, followed by the results obtained using the A-Ex approach with varying amounts of sociodemographic context included in the model prompt.

5.1. A-Ex vs. Ex-A

Model	Not Offensive			Offensive			Macro-avg		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Deepseek A-Ex	0.846	0.202	0.326	0.575	0.772	0.659	0.710	0.487	0.492
Deepseek Ex-A	0.872	0.228	0.361	0.588	0.891	0.708	0.730	0.559	0.535
Gemma A-Ex	0.892	0.232	0.368	0.606	0.909	0.727	0.749	0.570	0.548
Gemma Ex-A	0.889	0.267	0.411	0.604	0.553	0.557	0.746	0.410	0.494
Qwen A-Ex	0.848	0.601	0.703	0.704	0.899	0.790	0.776	0.750	0.747
Qwen Ex-A	0.881	0.574	0.695	0.697	0.925	0.795	0.789	0.749	0.745
Llama A-Ex	0.820	0.704	0.757	0.753	0.854	0.800	0.786	0.779	**0.779**
Llama Ex-A	0.806	0.684	0.740	0.739	0.844	0.788	0.772	0.764	0.764
Ministral A-Ex	0.950	0.270	0.421	0.589	0.987	0.737	0.770	0.628	0.579
Ministral Ex-A	0.903	0.435	0.587	0.642	0.956	0.768	0.772	0.695	0.677

Table 1
Results for the A-Ex and Ex-A approaches for the different models.

Table 1 shows precision, recall and F1 scores for the models that are prompted without sociodemographic information using the A-Ex and Ex-A approaches [3]. The last column shows the macro-averaged F1 scores, with the highest one for each model underlined and the overall highest score in bold. For Gemma, Qwen and Llama, the A-Ex approach performs better than the Ex-A one, while the opposite holds for Deepseek and Ministral. Additionally, Llama performs best across both approaches and with A-Ex it achieves the highest F1 score overall. For these reasons, we decided to focus on the A-Ex approach over the Ex-A one when investigating the effect of providing sociodemographic information.

When focusing on class-specific performance, most models achieve high recall but lower precision on the offensive class, except Gemma Ex-A where recall is also relatively low. For the Not Offensive class, precision is generally higher while recall remains lower. This is particularly true for Deepseek and Gemma with both approaches, and Ministral with A-Ex. To better understand this behavior, Figure 2

[3]To handle missing values, that occur when the generative model does not produce an answer that follows the prespecified format, these are treated as incorrect predictions and influence the recall on the corresponding class.

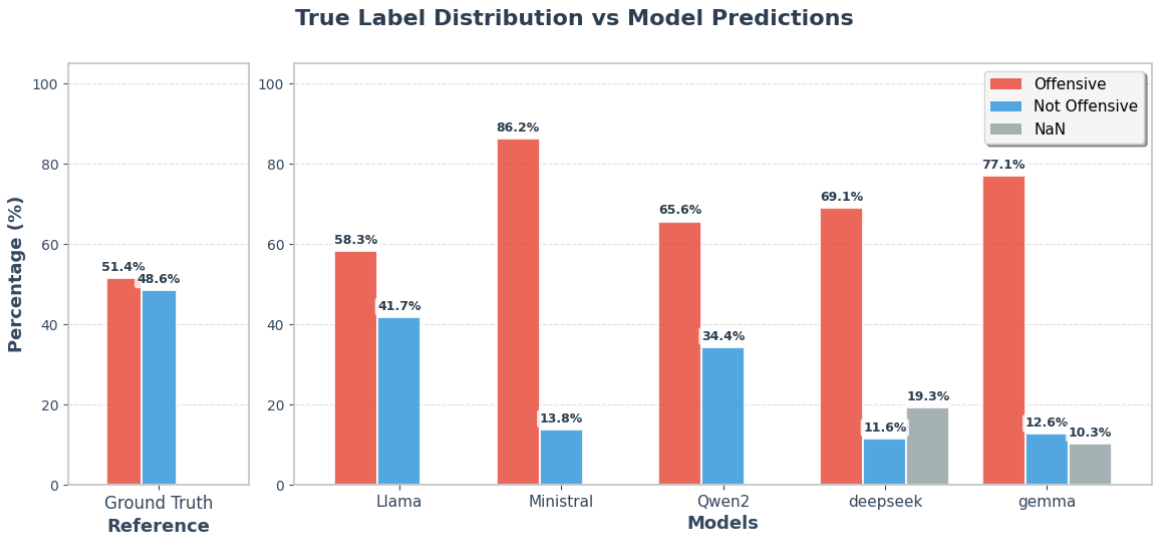

Figure 2: Individual annotator ground truth and predicted labels with the A-Ex approach.

compares the true label distribution with the predictions generated by each model. The plots reveal a strong tendency across models to over-predict the Offensive label, suggesting a risk of over-moderation which is further analyzed in Section 6.2.1. This effect is especially pronounced for Ministral, Deepseek, and Gemma, where the number of Not Offensive predictions is approximately a quarter of what is present in the ground-truth distribution. This agrees with the observed low recall on the Not Offensive class for these three models. Also note that Deepseek and Gemma in some cases fail to produce valid answers (e.g., generating both Offensive and Not Offensive labels simultaneously or producing excessively long responses), shown in grey in the bar chart.

5.2. Sociodemographic information

The results in Table 2 contain the precision, recall and F1 scores for all models using various combinations of sociodemographic traits along with the baseline where no sociodemographic traits are used, which is highlighted in grey. The final column shows macro-averaged F1 scores, with the highest one for each model underlined and the overall best in bold. The best performing model remains Llama across all combinations of sociodemographic variables, while the highest F1 score overall is obtained with Llama using both race and politics as sociodemographic variables.

For two out of the five models, namely Ministral and Deepseek, the baseline achieves the highest F1 score, with substantial decreases observed when sociodemographic variables are added. In contrast, for the remaining three models, incorporating sociodemographic variables results in a modest improvement in F1 score (ranging from 0.03-0.11). Specifically, for Gemma, including only race slightly outperforms the baseline (McNemar p-value = 0.043), for Qwen, combining gender and politics there is also a slight improvement over the baseline (McNemar p-value = 0.106), and for Llama, using race and politics a similar behavior is observed (McNemar p-value = 0.018). However, none of these improvements remain significant after applying a Bonferroni correction[4] for three tests (resulting in an adjusted $\alpha_{adj} = 0.017$ with a significance level $\alpha = 0.05$).

When examining class-specific performance, the same conclusions as in the overall results hold, with lower recall on the Not Offensive class and lower precision on the offensive one across all options. No

[4]The Bonferroni method controls for inflated Type I error when conducting multiple comparisons by dividing the overall significance level by the number of tests.

Model	Not Offensive			Offensive			Macro-avg		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Deepseek Baseline	0.846	0.202	0.326	0.575	0.772	0.659	0.710	0.487	0.584
Deepseek gender	0.856	0.195	0.318	0.578	0.757	0.655	0.717	0.476	0.487
Deepseek race	0.858	0.150	0.256	0.564	0.747	0.643	0.711	0.448	0.449
Deepseek politics	0.875	0.215	0.249	0.563	0.694	0.621	0.719	0.420	0.435
Deepseek gender race	0.854	0.176	0.292	0.572	0.768	0.655	0.713	0.472	0.474
Deepseek gender politics	0.870	0.151	0.257	0.567	0.706	0.629	0.723	0.428	0.443
Deepseek race politics	0.898	0.129	0.226	0.556	0.723	0.629	0.727	0.426	0.428
Deepseek gender race politics	0.866	0.147	0.251	0.557	0.718	0.627	0.711	0.433	0.505
Gemma Baseline	0.892	0.232	0.368	0.606	0.909	0.727	0.749	0.570	0.548
Gemma gender	0.898	0.247	0.388	0.607	0.906	0.727	0.752	0.577	0.557
Gemma race	0.884	0.245	0.383	0.615	0.915	0.736	0.749	0.580	0.559
Gemma politics	0.894	0.243	0.382	0.616	0.911	0.735	0.755	0.557	0.558
Gemma gender race	0.89	0.243	0.382	0.607	0.897	0.724	0.749	0.570	0.553
Gemma gender politics	0.892	0.237	0.375	0.611	0.899	0.727	0.751	0.568	0.551
Gemma race politics	0.889	0.236	0.373	0.616	0.903	0.732	0.752	0.57	0.553
Gemma gender race politics	0.878	0.231	0.365	0.614	0.891	0.727	0.746	0.561	0.546
Qwen Baseline	0.848	0.601	0.703	0.704	0.899	0.790	0.776	0.750	0.747
Qwen gender	0.843	0.602	0.702	0.704	0.894	0.788	0.774	0.748	0.745
Qwen race	0.853	0.601	0.705	0.705	0.902	0.792	0.779	0.752	0.748
Qwen politics	0.851	0.599	0.703	0.704	0.901	0.790	0.777	0.750	0.747
Qwen gender race	0.851	0.600	0.704	0.704	0.901	0.791	0.778	0.750	0.747
Qwen gender politics	0.847	0.611	0.710	0.709	0.896	0.791	0.778	0.753	0.750
Qwen race politics	0.855	0.601	0.706	0.706	0.904	0.792	0.780	0.752	0.749
Qwen gender race politics	0.849	0.603	0.705	0.706	0.899	0.791	0.777	0.751	0.748
Llama Baseline	0.820	0.704	0.757	0.753	0.854	0.800	0.786	0.779	0.779
Llama gender	0.817	0.702	0.755	0.751	0.851	0.798	0.784	0.777	0.777
Llama race	0.805	0.740	0.776	0.777	0.828	0.802	0.791	0.789	0.689
Llama politics	0.822	0.694	0.752	0.748	0.858	0.799	0.785	0.776	0.776
Llama gender race	0.817	0.719	0.650	0.762	0.847	0.802	0.789	0.783	0.784
Llama gender politics	0.816	0.690	0.748	0.745	0.853	0.795	0.78	0.771	0.771
Llama race politics	0.812	0.735	0.772	0.770	0.839	0.803	0.791	0.787	**0.788**
Llama gender race politics	0.815	0.708	0.758	0.755	0.849	0.799	0.785	0.778	0.778
Ministral Baseline	0.950	0.270	0.421	0.589	0.987	0.737	0.770	0.628	0.579
Ministral gender	0.959	0.190	0.317	0.565	0.992	0.72	0.762	0.591	0.518
Ministral race	0.958	0.199	0.330	0.567	0.992	0.722	0.763	0.595	0.526
Ministral politics	0.955	0.187	0.313	0.564	0.992	0.719	0.760	0.589	0.516
Ministral gender race	0.953	0.202	0.334	0.568	0.991	0.722	0.760	0.596	0.528
Ministral gender politics	0.957	0.193	0.321	0.565	0.992	0.720	0.761	0.592	0.521
Ministral race politics	0.960	0.194	0.323	0.566	0.992	0.721	0.763	0.593	0.522
Ministral gender race politics	0.951	0.196	0.325	0.566	0.991	0.720	0.759	0.593	0.522

Table 2
Results using combinations of the sociodemographic traits and baseline (A-Ex).

combination of sociodemographic traits consistently improves the model performance, neither overall nor on either of the two classes.

Overall, it appears that only small and not statistically significant improvements can be achieved with

the inclusion of sociodemographic variables for this task, while in some cases their inclusion causes large drops in performance. Another thing to note is that results do not appear to change much when comparing the baseline to the model prompted with sociodemographic information. Whether the model is capable of introducing some variability on this classification task based on the sociodemographic variables is further analyzed in Section 6.1

6. Analysis

In this section, we conduct additional analyses to address the questions raised, especially focusing on the best-performing model, namely Llama with race and political sociodemographic traits provided in the prompt. We analyzed the extent to which the models change their predictions when sociodemographics are included in the prompt and the effect of textual variables on the model performance. Finally, a qualitative analysis was conducted, with a focus on over-moderation and the explanations produced by the model.

6.1. Sensitivity of predictions to demographic traits

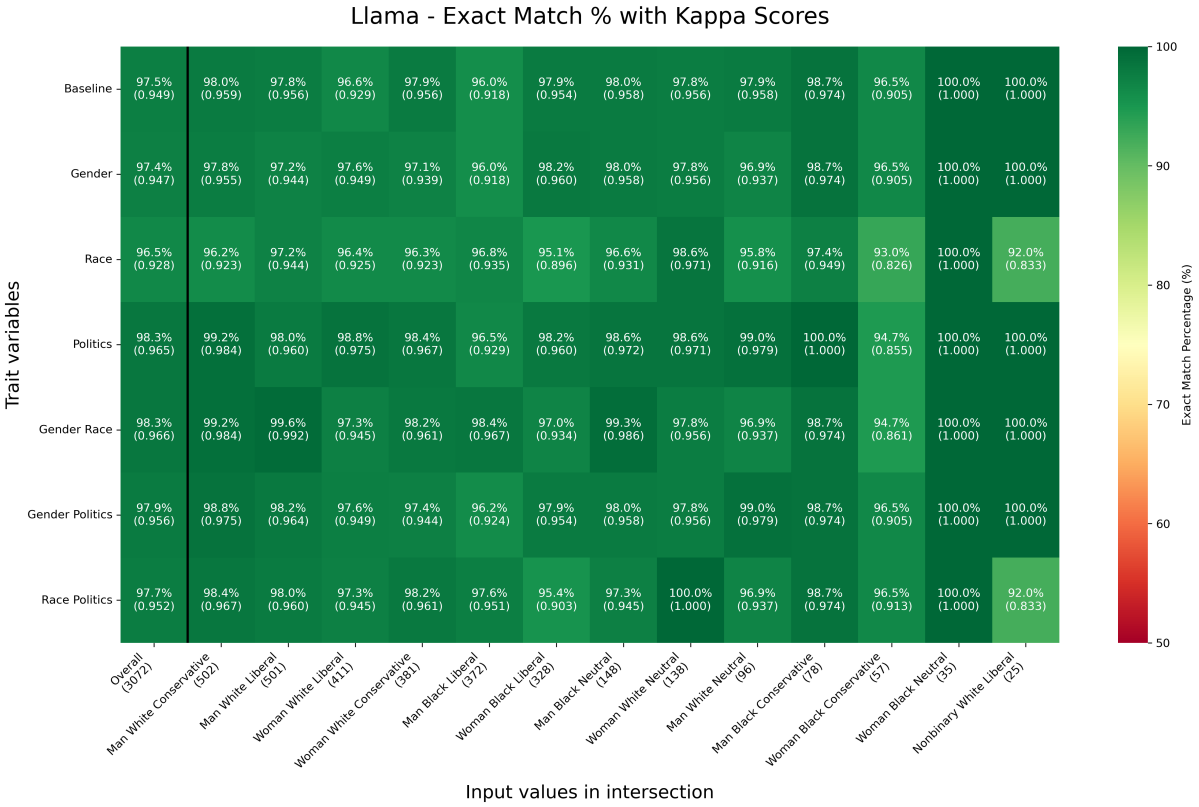

Figure 3: Measuring agreement between Llama's predictions with intersectional approach and any subset of sociodemographic variables.

To investigate whether the models change their labeling attribution depending on the sociodemographic traits provided, we compared the predictions made with the full set of traits against those made with a reduced set. We report both the percentage of exact matches (i.e., cases where the same label is assigned) and Cohen's Kappa scores. Since the models most often predict the label Offensive, the values of Cohen's Kappa can be distorted [40], so we comment based on the exact match percentage. Since some models produced NaN values, the instances where this occurs are removed for this analysis. To be more specific, when working with Deepseek, there were 32.7% of instances where some of the prompting methods led to missing values, while this occurred for 28.1% of instances with Gemma. We

further analyze results by looking at specific sociodemographic groups. Figure 3 contains the results of this investigation for Llama, which is the overall best performing model, while Figure 8 (Appendix C) contains the same plot for Deepseek, Qwen, Gemma and Ministral. The percentages of label matches are reported, with Cohen's Kappa values shown in brackets.

The first column in the heatmap compares predictions from the intersectional model (using all sociodemographic traits) against those from models prompted with reduced sets of traits, indicated on the y-axis, across the full dataset. The remaining columns analyze this in more detail by only retaining the instances which are annotated by individuals with the sociodemographic traits shown on the x-axis. For example, the entry in the second row, first column compares predictions from the intersectional model and the gender-only model across the entire dataset. The entry in the last row, second column compares model's predictions for the intersectional and the race-politics model only on the instances annotated by white conservative men annotators. This means that the intersectional model includes *Man White Conservative* in the prompt, while the race-politics model includes *White Conservative*. The idea is that if this number is small and there is low agreement, then the presence of gender *Man* influences the model predictions. Finally, note that results for some groups (e.g., *Nonbinary White Liberals* or *Black Women Neutral*) are based on small sample sizes (reported in brackets on the x-axis), and thus no strong conclusions should be drawn for these cases.

The results indicate that **all models tend to predict the same label regardless of sociodemographic traits**. The baseline and the intersectional model produce the same label on 97.5% of instances for Llama, with similar results for the other models. Similarly, when comparing intersectional model predictions with those from any other sociodemographic subset on the full dataset, they consistently match on over 96.5% of instances for Llama, with similar conclusions also when looking at the results for the other models.

Focusing on specific sociodemographic groups, the intersectional model does not differ drastically from the baseline or any other model with a subset of sociodemographic traits. The largest disagreement appears for the *Woman Black Conservative* group when comparing the intersectional model to the race-only model (including only *Black* in the prompt) with 93% of instances matching. Similar comments for the other models analyzed are reported in Appendix C.

6.1.1. Impact of textual variables on classification performance

We analyze the influence of the textual variables identified in the original dataset to see how they influence the model performance. Here we focus on vulgar language, since it resulted to be the most impactful on the model performance. Results on the other two text characteristics (African American English and anti-black language) are reported in Appendix C. The analysis is conducted on the best-performing model.

Vulgar	Class	Precision	Recall	F1-score	Support
True	Not Offensive	0.728	0.575	0.642	600
	Offensive	0.783	0.877	0.828	1051
False	Not Offensive	0.857	0.843	0.850	892
	Offensive	0.743	0.764	0.753	529

Table 3
Performance of Llama with race and political leaning sociodemographics on Offensive and Not Offensive classes on vulgar and non-vulgar texts.

Table 3 contains the precision, recall and F1 scores for both classes, with the first two rows for text instances containing vulgar terms and the last two for texts which do not. We note that the model performance on the negative class is much lower for vulgar texts than for non-vulgar ones, mainly caused by a low recall, which causes a drop of approximately 0.2 in F1 compared to that on non-vulgar texts. This means that the model struggles to identify instances that are not offensive and contain

vulgar terms, as can be seen from the low recall, which is caused by a large number of false positives on this split.

Overall, considering also the results in Table 6 and Table 7 (Appendix C), it appears that the presence of any of the three variables causes an increasing number of false positives, which points to over-moderation being a key issue.

6.2. Qualitative analysis

To better understand how the leveraged models moderate their outputs with respect to demographic information, we conducted a qualitative analysis. First, using the annotated dataset, we examined the phenomenon of over-moderation, focusing on instances where the model labeled inputs as Offensive, while human annotators did not (i.e. false positives). Second, we analyzed cases where the model failed to flag content human annotators considered Offensive (i.e. false negatives). Out of a total of $3,094$ annotations, there were 395 false positives (168 texts) and 254 false negatives (146 texts). Given the large number of experimental configurations and the time-intensive nature of manual evaluation, we focused our analysis only on the model with the highest performance, namely Llama with race and political leaning sociodemographics. Finally, we investigated whether the model's explanations varied based on the demographic information provided as input on the various Llama configurations.

6.2.1. Overmoderation

The models appear to struggle most with accurately categorizing sentences that contain vulgar language, as discussed in Section 6.1.1, especially slurs (derogatory terms targeting specific communities). It shows a tendency to label such sentences as Offensive, even when the slurs are used in non-offensive contexts (e.g., as interjections, in reappropriative usage, or in a neutral manner).

Taking a look at the sentences in Example 1 and Example 2:

(1) [Tweet] *to all my beautiful bitches: i love you all*

[Output] *the sentence is offensive because it uses derogatory language to address women*

(2) [Tweet] *I hate being called a bitch*

[Output] *the sentence is offensive because it conveys a negative emotion towards a gendered slur*

Llama misinterpreted the contextual use of the term "bitch", which in both cases is not directed at a specific person. In the first example, the word is used colloquially to express affection toward a specific group of people, while in the second, the speaker reflected on how being called the term makes them feel. In both instances, the model failed to capture the speaker's intent, resulting in an incorrect label assignment. Similarly, in the sentence: "I hate being called a fag", which was flagged as Offensive with the explanation: *the sentence is offensive because it contains derogatory language towards the lgbtq+ community*, again misinterpreting the polarity of the sentence.

The model tends to classify as offensive also those examples in which slurs are used to express more articulated opinions about the use of such terms, for instance:

(3) [Tweet] *'nig*as' should be used by Blacks only*

[Output] *the sentence is offensive because it promotes the use of a racial slur and implies that only black people should use it*

(4) [Tweet] *I am a transsexual. I am also a transgender person. And a trans person too! But don't you dare to call me a tran*y, cause you'll be in trouble. Especially if you're a bitchy b3nder.*

[Output] *the sentence is offensive because it contains derogatory language and a threat*

Such outputs can result in particular problems, as they risk censoring expressions of identity and discourse within marginalized and underrepresented communities [41].

Additionally, excessive moderation leads to losing the ability to deal with derogatory words, even when they are not directed at humans:

Correct prediction	Offensive	Not Offensive	Total
Both	1,311	1,040	2,351
Neither	216	385	601
Baseline	38	10	48
Race politics	15	57	72

Table 4
Comparison of correct predictions for Llama baseline and race politics.

(5) [Tweet] *my basement is full of those nasty cockroaches*

[Output] *the sentence is offensive because it contains derogatory language towards a group of insects*

In summary, **the model appears to over-moderate content, often flagging non-offensive posts as offensive**, particularly when vulgar terms are present. This suggests that it struggles to interpret such terms in context and may lack the pragmatic and sociocultural knowledge that human annotators typically rely on to distinguish between genuinely offensive language and reclaimed, informal, or ironic uses of such expressions.

6.2.2. Demographics can be misleading

When examining cases where the model labeled an input as Not Offensive but human annotators judged it Offensive, we found that the sentences were often ambiguous and tended to divide annotators as well. Since the model is asked to examine the same example multiple times with a slight difference in the identity to impersonate, the possibility that a prediction would be correct for one group of annotators and incorrect for another was expected. For this reason, we focused on the texts where all the annotators agreed on labeling them as Offensive (8 out of 146). Interestingly, we noticed that the model prompted with demographics tended to justify the offensive content as simple opinions and discussions, especially focusing on the absence of hate speech and explicit language. On the other hand, when looking at the output of the baselines, in 5 cases it actually provided the correct label and explanation, as in the following examples:

(6) [Tweet] *america is trash, too many immigrants that is why*

[Baseline output] *the sentence is offensive because it conveys a xenophobic message*

[Race-politics output] *the sentence is not offensive because it is a political statement that expresses a negative opinion about a country and its immigration policies*

(7) [Tweet] *The way these nig*as hate I hope they get a check for it*

[Baseline output] *the sentence is offensive because it contains a racial slur and promotes hate*

[Race-politics output] *the sentence is not offensive because it is a phrase used to express frustration and does not contain hate speech*

Following a more systematic comparison between the two models, it is possible to notice that although the demographic-based model generally produces more correct labels than the baseline, this improvement is driven primarily by predictions of the Not Offensive class (Table 4). The number of such cases is not large, but it highlights the **importance of conducting a careful error analysis to assess the risk of demographic features being misleading, particularly when influencing predictions within a single class**.

6.2.3. Explanations

To analyze the variation of the explanations among the different configurations of Llama compared to its baseline, we vectorized only the strict explanation, removing the more stable portion of the response. Thus, we retained what comes after `[the sentence is offensive/is not offensive][because]`. We used BERT-base-uncased[5] to generate 768-dimensional vector representations for the baseline and each configuration individually. We adopted k-means clustering with the number of clusters chosen to maximize the Silhouette Coefficient [42], which measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation), and is bounded between -1 and $+1$.

When computing the similarity across the full annotated dataset, the obtained clusters were grouped based on whether the model classified a text as Offensive or Not Offensive, regardless of whether the explanation was generated by the baseline or by the demographic-based model (Section D). This demonstrates that the explanations, as well as the label distribution, did not present a strong variation when demographics were provided.

In a second step, we filtered explanations based on the assigned Offensive or Not Offensive label to examine whether specific patterns of the annotated texts were reflected in the explanations. In these cases, the clusters tend to be even more sparse, remaining homogeneous despite the textual characteristics, whether vulgar language, African American English (AAE), or content targeting Black people.

Overall, the Silhouette Coefficient is close to random in all settings (Table 5), suggesting the **absence of a clear clustering tendency**, thus a strong similarity between explanations generated by the baseline and the demographic-based models.

Model	Overall	Offensive	Not Offensive
gender	0.150	0.121	0.132
race	0.158	0.128	0.141
politics	0.152	0.113	0.145
gender race	0.152	0.107	0.134
gender politics	0.147	0.119	0.141
race politics	0.150	0.120	0.143
gender race politics	0.148	0.106	0.140

Table 5
Silhouette Coefficient of clusters obtained comparing explanations from the Llama baseline with each of the settings considering the dataset in its completeness (Overall), and filtered based on texts predicted as Offensive and Not Offensive.

7. Conclusion and Future Works

This work analyzed the effect of sociodemographic prompting on small generative LLMs challenged on a highly subjective linguistic phenomenon and a complex task. We focused not only on the effect of these traits on model performance, but also on whether they had an impact on labeling behavior and the generation of explanations. Our analysis employed a combination of quantitative and qualitative methods, also considering the role of three textual variables.

To answer RQ1, we found that introducing sociodemographic traits, either individually or in intersection, did not lead to consistent improvements over the baseline.

For RQ2, we observed that labeling behavior remained highly stable across conditions, with predictions agreeing on more than 95% of cases regardless of the traits included in the prompt. As such, no particular trait appeared to be influential, with only small deviations on specific sociodemographic groups which were not consistent across models. This contrasts with the findings presented by the dataset authors,

[5] https://huggingface.co/google-bert/bert-base-uncased

who observed a correlation between labeling behavior in texts containing AAE and political leaning. This is not reflected by the LLM annotations, which have little variation regardless of whether the political leaning is specified in the prompt or not.

Finally, for RQ3, the additional analyses of the explanations also showed no variability induced by the sociodemographic prompting. Instead, they tended to group according to the output label. Although the explanations remained consistent across different traits, they provided valuable insights into the model's decision-making process and over-moderation behavior.

Our results show that **small models struggle to exploit sociodemographic information on a sensitive task such as offensive speech detection, both at the label and explanation level**. Moreover, this work further highlights the importance of thorough analysis when this prompting technique is used. Specifically, the study demonstrates the importance of going beyond performance metrics, consistently investigating models' behaviors, and deeply questioning their reliability.

In the future, we intend to combine sociodemographic prompting with reasoning (CoT) models and approaches, which may encourage the model to reflect different perspectives more effectively. This would be an extension of the Ex-A approach, which was tried in the baseline case but not with sociodemographic traits. Another possibility would be to adopt group-specific few-shot strategies, where examples and explanations come from individuals of the relevant demographic group rather than from consensus cases, to steer the model towards various perspectives.

Limitation

We identify and briefly describe the major limitations of this work.

Models Scale The experiments are conducted using relatively small LLMs (7B and 8B parameters) due to cost and computational resource constraints. Including results from larger models would allow for a more comprehensive evaluation and help determine whether more capable systems are better able to leverage sociodemographic information.

Single dataset Our analyses were carried out on a single dataset, chosen because of the extensive quantitative and qualitative evaluation it required and the associated computational limitations. Extending the study to additional datasets that capture subjective and disaggregated phenomena would be valuable to assess whether the findings generalize and whether sociodemographic information can play a meaningful role in broader contexts. Moreover, a focus on datasets with strong sociodemographic impact on a phenomenon that makes the model less prone to over-moderation would help better assess whether the LLMs can modulate their predictions based on those.

Prompt variations While we experimented with some prompts for the task and ways to specify the sociodemographic information, our exploration was not exhaustive. Different prompt formulations, demographics specification or ordering of sociodemographic attributes may influence model behavior in ways we did not capture. As such, our results may partly reflect the specific prompt choices we adopted rather than generalizable model tendencies.

Clustering Sentence Embeddings We employed simple techniques to obtain sentence embeddings and clustering them, to extract whether signals of the inclusion of sociodemographic traits could be picked up on either positive or negative explanations. Specifically, we relied on BERT embeddings with k-means clustering, which is straightforward but comes with limitations. The first is that since we use a general-purpose model, sentences describing similar phenomena may be embedded very closely. An additional limitation is the k-means assumption of spherical clusters. Hence, this approach should be viewed more as an exploration to assess whether strong signals could be identified.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120.

[2] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, Annollm: Making large language models to be better crowdsourced annotators, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), 2024, pp. 165–190.

[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. URL: https://aclanthology.org/2021.bppf-1.3/. doi:10.18653/v1/2021.bppf-1.3.

[4] T. Horych, C. Mandl, T. Ruas, A. Greiner-Petter, B. Gipp, A. Aizawa, T. Spinde, The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 1370–1386. URL: https://aclanthology.org/2025.findings-naacl.75/. doi:10.18653/v1/2025.findings-naacl.75.

[5] X. Wang, H. Kim, S. Rahman, K. Mitra, Z. Miao, Human-llm collaborative annotation through effective verification of llm labels, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–21.

[6] S. Santy, J. Liang, R. Le Bras, K. Reinecke, M. Sap, Nlpositionality: Characterizing design biases of datasets and models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 9080–9102.

[7] 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, Language Resources and Evaluation (2024) 1–28.

[8] P. Fortuna, M. Dominguez, L. Wanner, Z. Talat, Directions for NLP practices applied to online hate speech detection, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 11794–11805. URL: https://aclanthology.org/2022.emnlp-main.809/. doi:10.18653/v1/2022.emnlp-main.809.

[9] V. Prabhakaran, A. Mostafazadeh Davani, M. Diaz, On releasing annotator-level labels and information in datasets, in: C. Bonial, N. Xue (Eds.), Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 133–138. URL: https://aclanthology.org/2021.law-1.14/. doi:10.18653/v1/2021.law-1.14.

[10] H. Al Kuwatly, M. Wich, G. Groh, Identifying and measuring annotator bias based on annotators' demographic characteristics, in: S. Akiwowo, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 184–190. URL: https://aclanthology.org/2020.alw-1.21/. doi:10.18653/v1/2020.alw-1.21.

[11] 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. URL: https://aclanthology.org/2023.emnlp-main.212/. doi:10.18653/v1/2023.emnlp-main.212.

[12] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, T. Hashimoto, Whose opinions do language models reflect?, in: International Conference on Machine Learning, PMLR, 2023, pp. 29971–30004.

[13] E. Hwang, B. Majumder, N. Tandon, Aligning language models to user opinions, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5906–5919. URL: https://aclanthology.org/2023.findings-emnlp.393/. doi:10.18653/v1/2023.findings-emnlp.393.

[14] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using language models to simulate human samples, Political Analysis 31 (2023) 337–351. doi:10.1017/pan.2023.2.

[15] A. Wang, J. Morgenstern, J. P. Dickerson, Large language models that replace human participants can harmfully misportray and flatten identity groups, Nature Machine Intelligence (2025) 1–12.

[16] T. Beck, H. Schuff, A. Lauscher, I. Gurevych, How (not) to use sociodemographic information for subjective nlp tasks, arXiv preprint arXiv:2309.07034 (2023).

[17] T. Beck, H. Schuff, A. Lauscher, I. Gurevych, Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 2589–2615.

[18] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, N. A. Smith, Annotators with attitudes: How annotator beliefs and identities bias toxic language detection, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5884–5906. URL: https://aclanthology.org/2022.naacl-main.431/. doi:10.18653/v1/2022.naacl-main.431.

[19] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5477–5490.

[20] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, S.-Y. Yun, Hare: Explainable hate speech detection with step-by-step reasoning, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 5490–5505.

[21] J. Hessel, A. Marasovic, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, Y. Choi, Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest, 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. 688–714. URL: https://aclanthology.org/2023.acl-long.41/. doi:10.18653/v1/2023.acl-long.41.

[22] E. Hwang, V. Shwartz, MemeCap: A dataset for captioning and interpreting memes, 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. 1433–1445. URL: https://aclanthology.org/2023.emnlp-main.89/. doi:10.18653/v1/2023.emnlp-main.89.

[23] 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. URL: https://aclanthology.org/2020.acl-main.485/. doi:10.18653/v1/2020.acl-main.485.

[24] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[25] J. Schäfer, A. Combs, C. Bagdon, J. Li, N. Probol, L. Greschner, S. Papay, Y. M. Resendiz, A. Velutharambath, A. Wührl, et al., Which demographics do llms default to during annotation?, arXiv preprint arXiv:2410.08820 (2024).
[26] A. C. Curry, G. Attanasio, Z. Talat, D. Hovy, Classist tools: Social class correlates with performance in nlp, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 12643–12655.
[27] A. C. Curry, Z. Talat, D. Hovy, Impoverished language technology: The lack of (social) class in nlp, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 8675–8682.
[28] F. Plaza-del Arco, A. Curry, S. Paoli, A. C. Curry, D. Hovy, Divine llamas: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 4346–4366.
[29] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, et al., Towards measuring the representation of subjective global opinions in language models, arXiv preprint arXiv:2306.16388 (2023).
[30] T. Giorgi, L. Cima, T. Fagni, M. Avvenuti, S. Cresci, Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 19, 2025, pp. 653–670.
[31] M. Orlikowski, J. Pei, P. Röttger, P. Cimiano, D. Jurgens, D. Hovy, Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 2092–2111. URL: https://aclanthology.org/2025.acl-long.104/. doi:10.18653/v1/2025.acl-long.104.
[32] H. Sun, J. Pei, M. Choi, D. Jurgens, Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 2025, pp. 845–854.
[33] T. Hu, N. Collier, Quantifying the persona effect in LLM simulations, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10289–10307. URL: https://aclanthology.org/2024.acl-long.554/. doi:10.18653/v1/2024.acl-long.554.
[34] P. F. Balestrucci, M. Oliverio, S. M. Lo, L. Anselma, V. Basile, C. Bosco, A. Mazzei, V. Patti, When figures speak with irony: Investigating the role of rhetorical figures in irony generation with llms (2025).
[35] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al., Deepseek llm: Scaling open-source language models with longtermism, arXiv preprint arXiv:2401.02954 (2024).
[36] Meta, Meta-llama-3.1-8b-instruct, 2024. URL: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.
[37] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, Z. Fan, Qwen2 technical report, 2024. URL: https://arxiv.org/abs/2407.10671. arXiv:2407.10671.
[38] M. AI, Ministral-8b-instruct-2410, 2025. URL: https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.
[39] Google, Gemma: Introducing new state-of-the-art open models, 2024. URL: https://blog.google/technology/developers/gemma-open-models/.
[40] A. R. Feinstein, D. V. Cicchetti, High agreement but low kappa: I. the problems of two paradoxes,

Journal of Clinical Epidemiology 43 (1990) 543–549. URL: https://www.sciencedirect.com/science/article/pii/089543569090158L. doi:https://doi.org/10.1016/0895-4356(90)90158-L.
[41] L. Draetta, C. Ferrando, M. Cuccarini, L. James, V. Patti, ReCLAIM project: Exploring Italian slurs reappropriation with large language models, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 335–342. URL: https://aclanthology.org/2024.clicit-1.40/.
[42] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

A. Information about the dataset

The annotators had to answer the questions:

- How toxic/hateful/disrespectful or offensive does this post seem to you?
- How toxic/hateful/disrespectful or offensive does this post seem to anyone?

The authors then took the average rating and obtained an *offensiveness* score [18]. The complete distribution of it is presented in Figure 4.

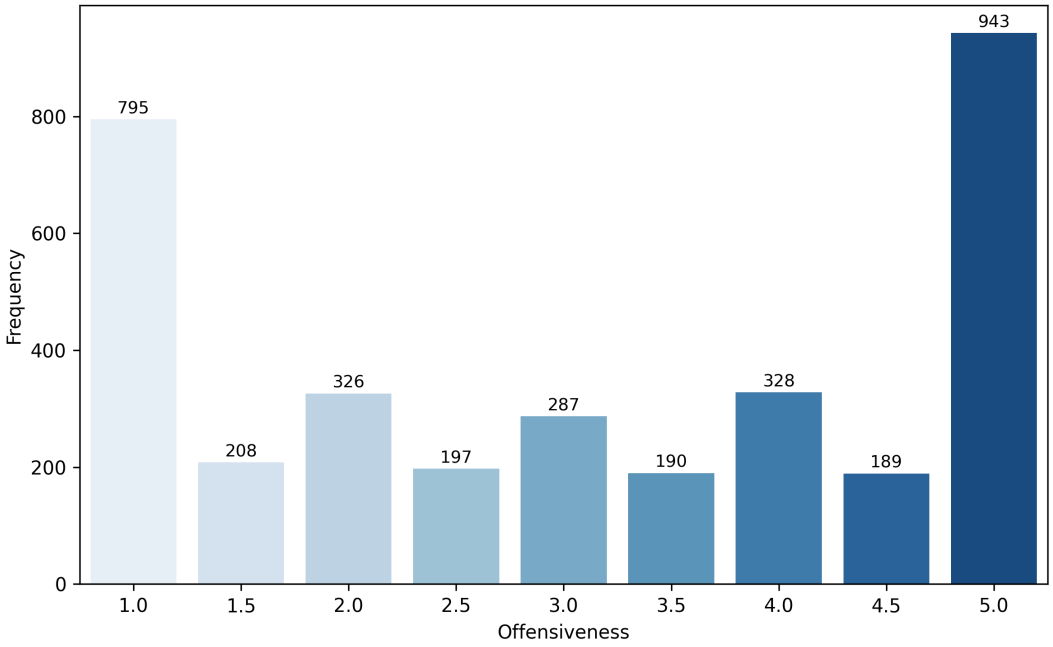

Figure 4: Distribution of the average offensiveness rating from the unfiltered dataset.

B. Prompt Details

B.1. Few-Shot Examples

Offensive Examples:

- "*A girl that looked like she worth the world to ya is around the corner getting fucked for a perk #Facts*"
- "*to call black people niggas is just funny*"

Not Offensive Examples:

- "*as a black person, i disagree with white people saying nigga*"
- "*bitch it's fucking spring, why is this weather wildin? :triumph:*"

B.2. Full Prompt (A-Ex)

```
You are a person asked to provide precise information about offensiveness of a
sentence.
### Instruction:
Given the sentence in input (Input) taken from a social network, your task is to
determine if the text is offensive and why. Some of the sentences may contain
hate or explicit language, please provide an answer in any case.

Provide in output (Output) a single label among the following arrays ['offensive',
'not offensive']. Add a brief explanation following the format [the sentence is
offensive/is not offensive] [because] [explanation].

Example 1:
Input: A girl that looked like she worth the world to ya is around the corner
getting fucked for a perk #Facts
Output: [The sentence is offensive][because][it conveys a stereotypical message
suggesting that women use their bodies to obtain benefits and contains vulgar
language]

Example 2:
Input: as a black person, i disagree with white people saying nigga
Output: [The sentence is not offensive][because][the person is expressing an
opinion about the misuse of the racial slur 'nigga']

Example 3:
Input: to call black people niggas is just funny
Output: [The sentence is offensive][because][it promotes the use of a racial
slur]

Example 4:
Input: bitch it's fucking spring, why is this weather wildin? :triumph: Output:
[The sentence is not offensive][because][it is just a colloquial way to comment
on the weather]

Example to label:
Input: Women are not fucking inferior to men
Output:
```

Figure 5: Example of full prompt for A-Ex approach

B.3. Full Prompt (Ex-A)

```
You are a person asked to provide precise information about offensiveness of a
sentence.
### Instruction:
Given the sentence in input (Input) taken from a social network, your task is to
determine if the text is offensive and why. Some of the sentences may contain
hate or explicit language, please provide an answer in any case.
Provide in output (Output) a single label among the following arrays ['offensive',
'not offensive']. Explain your reasoning first, and return a single label.
Follow this format [reasoning explanation] [so] [the sentence is offensive/is
not offensive].

Example 1:
Input: A girl that looked like she worth the world to ya is around the corner
getting fucked for a perk #Facts
Output: [The sentence conveys a stereotypical message suggesting that women use
their bodies to obtain benefits and contains vulgar language][so][the sentence
is offensive]

Example 2:
Input: as a black person, i disagree with white people saying nigga
Output: [The person is expressing an opinion about the misuse of the racial
slur 'nigga'] [so][the sentence is not offensive]

Example 3:
Input: to call black people niggas is just funny Output: [The sentence promotes
the use of a racial slur][so][the sentence is offensive]

Example 4:
Input: bitch it's fucking spring, why is this weather wildin? :triumph:
Output: [The sentence is just a colloquial way to comment on the weather][so]
[the sentence is not offensive]


Example to label:
Input: Women are not fucking inferior to men
Output:
```

Figure 6: Example of full prompt for Ex-A approach

B.4. Example Prompt for A-Ex approach with sociodemographics

```
You are a [[race]] [gender|person] [[with [politics] political leaning]] asked
to provide precise information about offensiveness of a sentence.
### Instruction:

Given the sentence in input (Input) taken from a social network, your
task is to determine if the text is offensive and why. Some of the sentences
may contain hate or explicit language, please provide an answer in any case.
Provide in output (Output) a single label among the following arrays ['offensive',
'not offensive']. Add a brief explanation following the format [the sentence is
offensive/is not offensive] [because] [explanation].

Example 1:
Input: A girl that looked like she worth the world to ya is around the corner
getting fucked for a perk #Facts
Output: [The sentence is offensive][because][it conveys a stereotypical message
suggesting that women use their bodies to obtain benefits and contains vulgar
language]

Example 2:
Input: as a black person, i disagree with white people saying nigga
Output: [The sentence is not offensive][because][the person is expressing an
opinion about the misuse of the racial slur 'nigga']

Example 3:
Input: to call black people niggas is just funny
Output: [The sentence is offensive][because][it promotes the use of a racial
slur]

Example 4:
Input: bitch it's fucking spring, why is this weather wildin? :triumph:
Output: [The sentence is not offensive][because][it is just a colloquial way to
comment on the weather]


Example to label:
Input: Women are not fucking inferior to men
Output:
```

Figure 7: Example of full prompt for A-Ex approach with sociodemographic information

C. Additional results

C.1. Sensitivity of predictions to demographic traits

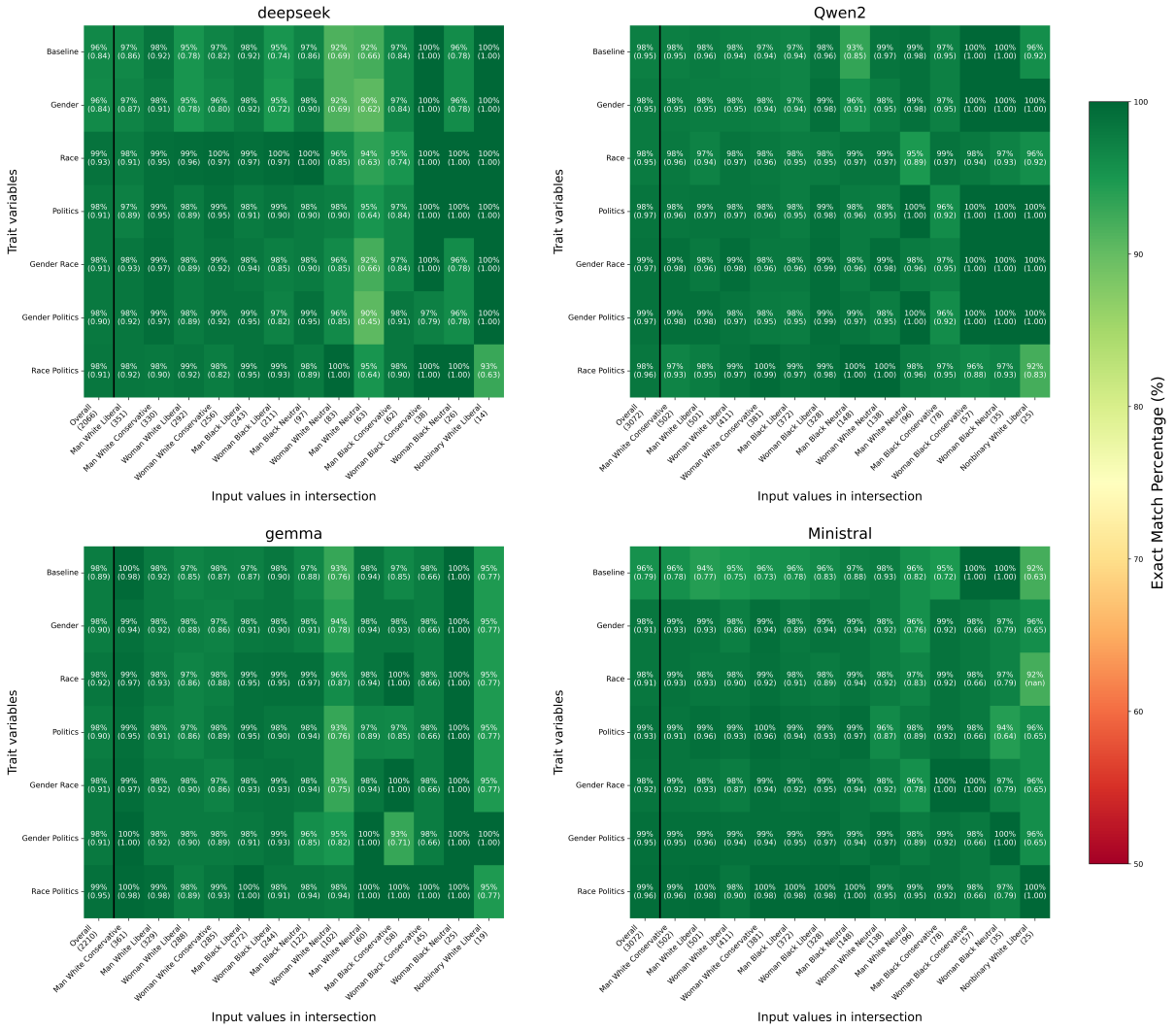

Figure 8: Measuring agreement between predictions with intersectional approach and any subset of socio-demographic variables for deepseek, Qwen2, gemma and Ministral

Here we briefly comment the results of the analysis conducted in Section 6.1 for the other models considered, shown in Figure 8.

For Deepseek, the highest variability occurs for "Man White Neutral" when comparing the intersectional model to the gender-and-politics model ("Man Neutral" in the prompt), with 90% of predictions matching. In Qwen, this happens for "Man Black Conservative" when comparing the intersectional model to the baseline, with agreement on 93% of instances. Gemma shows the most difference for "Woman White Neutral" when comparing the intersectional model with the baseline, the politics only model (i.e. "Neutral" included in the prompt) and the gender race model (i.e., "Woman White" included in the prompt), as well as for "Man Black Conservative" when comparing the intersectional model to gender politics (i.e., "Man Conservative" included in the prompt) with 93% of exact matches in all cases. Finally for Ministral this occurs for "Man White Liberal" compared to the baseline with 94% of exact matches, with all others being above 95%.

C.2. Impact of textual variables on classification performance

Here we report and briefly comment the results of the same analysis conducted in Section 6.1.1 on the other two textual variables of interest, namely whether the text contains AAE or targets Black People.

isAAE	Class	Precision	Recall	F1-score	Support
True	Not Offensive	0.853	0.732	0.788	973
	Offensive	0.642	0.792	0.709	591
False	Not Offensive	0.746	0.742	0.744	519
	Offensive	0.865	0.868	0.866	989

Table 6
Model performance on positive and negative classes on text containing or not AAE.

Table 6 contains the same metrics where the first two rows are for texts which contain AAE and the last two for those that do not. We observe that the model performs much worse in identifying offensive text when AAE is present, with an F1 score lower by approximately 0.15 points compared to those on texts which are not in AAE. This is caused by a low precision on this class due to a large number of False Positives in this case, showing a tendency of the model to classify text containing AAE as offensive even when this is not the case.

tBP	Class	Precision	Recall	F1-score	Support
True	Not Offensive	0.392	0.365	0.378	85
	Offensive	0.926	0.934	0.930	725
False	Not Offensive	0.838	0.758	0.796	1407
	Offensive	0.656	0.759	0.704	855

Table 7
Model performance on positive and negative classes on text which targets or not black people.

Table 7 contains model performance results on both classes, where the first two rows are for text targeting black people (tBP) and the last two for text that does not. On text targeting black people, the model performance is much lower on the negative class, with a low precision and low recall, caused by a large number of false positives. This means that the model struggles to identify correctly texts which target black people in a non-offensive way. For the positive class instead, the model performance is much higher on this split of the dataset.

D. Explanation analysis

We computed the similarity of the explanations between the baseline and each of the demographic-based settings on the full dataset to understand whether they separated based on the model or on the predicted label. Table 8 presents the composition of the obtained clusters. The last two columns detail how many members of each cluster are explanations generated by the baseline and the demographic-based models (Model comparison), and how many come from positive and negative predictions (Prediction comparison). Results show that the obtained clusters tend to be homogenous with respect to the source of the explanation, and strongly separated on their predicted label.

Traits	Clusters	Model comparison		Prediction comparison	
		Baseline	Dem-model	Not Offensive	Offensive
Gender	cluster 0	316	357	538	135
	cluster 1	415	480	103	792
Race	cluster 0	294	357	559	92
	cluster 1	437	510	113	834
Politics	cluster 0	416	484	109	791
	cluster 1	315	334	507	142
Gender race	cluster 0	415	591	118	888
	cluster 1	316	418	593	141
Gender politics	cluster 0	414	535	114	835
	cluster 1	317	397	556	158
Race politics	cluster 0	312	419	601	130
	cluster 1	419	566	137	848
Gender race politics	cluster 0	316	431	601	146
	cluster 1	415	630	132	913

Table 8
Cluster composition comparing baseline and sociodemographic models based on explanation source and predicted label.