HurtLens: A Perspectivist Corpus Analysis of Hurtful Language

Samuele D'Avenia, Eliana Di Palma, Marta Marchiori Manerba, Valerio Basile
Computer Science Department, University of Turin, Turin, Italy
{samuele.davenia, eliana.dipalma, marta.marchiorimanerba, valerio.basile}@unito.it

Abstract

Offensive language detection systems often rely on majority-aggregated annotations, overlooking the diversity of perspectives that shape how different communities perceive harm. In this contribution, we introduce HurtLens, a perspectivist corpus of hurtful language leveraging four disaggregated datasets which are automatically enriched through HurtLex lemmas, a multilingual resource of offensive and derogatory terms. Using mixed-effects modeling, we investigate how annotators' sociodemographic backgrounds, the presence of specific types of offensive language (through Hurtlex categories) and their interaction influence offensiveness ratings. Our analysis reveals that offensiveness ratings are influenced both by annotators' sociodemographic characteristics (particularly when considering them in intersection) and by the presence of specific types of offensive language. Additionally, we identify significant interaction effects showing that different demographic groups vary in their sensitivity to texts containing particular types of offensive language.

Keywords: offensive language, perspectivism, language resources, sociodemographic analysis

WARNING: *This paper contains examples of offensive or upsetting content.*

1. Introduction

A single word can hurt, but not everyone is offended by the same type of words in a specific context. Offensiveness is a complex phenomenon, where context and individual-subjectivity play a significant role (Kiritchenko et al., 2021).

The problem of offensive speech detection is not new in literature, neither in NLP nor in other related fields such as linguistics, social sciences, and communication sciences (Fortuna and Nunes, 2018). However, when it comes to offensive speech detection, what is mainly done is to annotate datasets based on the majority opinion of annotators, thus losing essential information such as minority views (Mostafazadeh Davani et al., 2022). A turning point in this regard has been provided by the perspectivist approach, which calls for the release of disaggregated resources and systems that learn from disagreement (Cabitza et al., 2023a).

In this work, we introduce HurtLens, a perspectivist corpus of hurtful language[1] that preserves disaggregated annotations across sociodemographic groups. This corpus enriches four existing resources under the lens of HurtLex lemmas, enabling the analysis of how different sociodemographic groups respond to different types of hurtful expressions. HurtLex is a multilingual lexicon containing offensive and hateful words, which are mapped to 17 categories indicating the semantic area of the word used to offend. For example "*wh*re*" falls in the category of words related to prostitution, while "*pig*" in that of animals (Bassignana et al., 2018).

Importantly, HurtLens includes both offensive and non-offensive uses of potentially hurtful words, reflecting the inherently contextual nature of offensiveness. Using our resource, we analyze how sociodemographic characteristics (*who*), the presence of specific types of offensiveness (*what*) and their interaction (*who reacts to what*) jointly shape perceived offensiveness. We articulate our work into three research questions:

RQ1 How do annotators' sociodemographics influence their offensiveness ratings of texts?

RQ2 How does the presence of lemmas belonging to specific HurtLex categories influence the offensiveness ratings?

RQ3 Do certain demographic groups exhibit different sensitivity to texts including certain HurtLex categories of offensive language?

Following previous works by Homan et al. (2024), we leverage *multilevel modelling* (Gelman and Hill, 2006) (or mixed-effects modelling), to analyze these three levels. This approach enables the examination of how sociodemographic factors, categories of hurtful language, and group-specific sensitivities interact, while accounting for inherent variability in both the text itself and the annotators.

Our analysis reveals that **intersectionality of sociodemographic traits** provides a more comprehensive explanation of rating behaviour compared to independent sociodemographics. Furthermore, we observe that the presence of specific types of hurtful words also informs the ratings pre-

[1] Throughout this work, we use the terms *hurtful* and *offensive* interchangeably to refer to language that may cause harm or be perceived as disrespectful, since both terms encompass a spectrum of harmful language phenomena (Poletto et al., 2020).

Proceedings of NLPerspectives @ LREC 2026, pages 44–55
12 May 2026.

diction. Finally, analyzing the interaction between sociodemographics and types of hurtful lemmas, **we uncover different sensitivities across *age* and *race* groups to different types of offensive lemmas.**

The full code is publicly available[2].

2. Related Works

In recent years, the growing presence of offensive and discriminatory language in public debate and on online platforms has attracted increasing attention, becoming a major concern for society and the scientific community in various fields. An approach to addressing this issue has been to develop models for the recognition of hate speech (Basile et al., 2019; Zampieri et al., 2020), requiring linguistic resources for model training and benchmarking mostly based on annotated texts taken from social media platforms (Poletto et al., 2020; Alkomah and Ma, 2022; Yu et al., 2024; Fortuna et al., 2020; Ollagnier, 2024).

Although previous research has focused mainly on textual resources, there are also studies that treat words as clues for analysing hate speech, using lexical knowledge to identify offensive language. Lexica based on this assumption are presented, for example, in Wiegand et al. (2018) and Bassignana et al. (2018). HateWiC (Wiegand et al., 2018) provides a basic and automatically expanded lexicon of words in context, based on the assumption that offensive terms constitute a subset of negatively polarized expressions. HurtLex (Bassignana et al., 2018), on the other hand, is a multilingual lexicon of hate originally developed for Italian and organised into 17 semantic categories. The lexicon was then expanded through links to synset-based lexical resources such as MultiWordNet and BabelNet, and extended to multiple languages through semi-automatic translation and expert annotations.

In the same years, a new paradigm in linguistic annotation has emerged, highlighting the inherently subjective nature of human annotation (Aroyo and Welty, 2015). In this context, disagreement is no longer seen as background noise, but as a meaningful signal (Uma et al., 2021; Plank, 2022). Building on this view, the perspectivist approach seeks to model and preserve annotators' viewpoints (Basile, 2021; Cabitza et al., 2023b). This has led to the spread of disaggregated corpora and datasets (Sap et al., 2022; Sachdeva et al., 2022; Frenda et al., 2023), and a shift in research towards the analysis of the perspectives that emerge from the annotation process itself (Mostafazadeh Davani et al., 2022; Homan et al., 2024; Sap et al., 2022).

[2] https://github.com/SDavenia/hurt_persp

Previous works have analyzed the effect of sociodemographic variables in subjective NLP tasks. Hu and Collier (2024) analyse the role of non-intersectional persona variables, finding that they explain up to 10% of the variability in those datasets. Homan et al. (2024) show that safety judgments in conversational AI are highly subjective and shaped by intersectional demographic factors, with Bayesian multilevel models revealing how gender, race/ethnicity, age, and education jointly influence annotators' perceptions of conversational harm and surface underrepresented perspectives.

3. HurtLens: A Corpus of Perspectivist Hurtfulness

In this section, we describe the construction of HurtLens, a corpus of social media posts featuring lemmas from the HurtLex lexicon of hurtful words (Bassignana et al., 2018), encompassing both offensive and non-offensive usages. We first present the lexicon and the source datasets upon which HurtLens relies, and then describe the construction process.

3.1. HurtLex

As previously mentioned, HurtLex (Bassignana et al., 2018) is a multilingual computational lexicon of offensive and hateful expressions, originally derived from the Italian lexicon *Le Parole per Ferire* developed by De Mauro (2016). We chose this lexical resource for its detailed categorical structure, which enables fine-grained analyses across different semantic types of offensive language. In addition, it is a widely adopted resource explicitly designed with a multilingual perspective (Stanković et al., 2020; Stamou et al., 2022; Tontodimamma et al., 2023; Osenova, 2024) and its effectiveness for offensive language detection has been demonstrated in numerous studies (Koufakou et al., 2020; Giordano and Di Buono, 2023; Árcos and Pérez, 2023).

In this work, we leverage the English HurtLex-core lexicon, containing 501 lemmas. Each lemma is associated with a semantic category tag that reflects the taxonomy defined in the original Italian lexicon and preserved across its multilingual extension. These categories capture the kind of harmful or sensitive concept the word expresses, with some referring to social groups (e.g. ethnic slurs), or others to personal characteristics (e.g. physical or cognitive disability).

One of the authors of this paper, with a Linguistics training, manually reviewed each entry and removed entries with no known offensive usage. Entries were retained if they exhibited at least one of the following: (i) explicit offensiveness or abusive

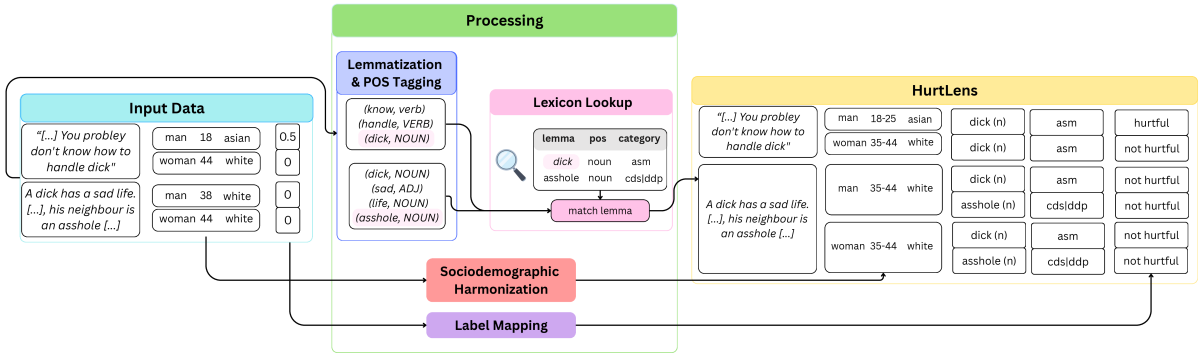

Figure 1: HurtLens construction steps. Example with entries from the SBIC dataset.

meaning (e.g., "*wh*re*"), (ii) non-literal derogatory usage arising from semantic shift or metaphor (e.g., "*pig*"), (iii) negative evaluative connotation without direct insult and (iv) vulgarity, such as terms related to sexuality that may occur in derogatory contexts.

Additionally, the same author reviewed the semantic categories to handle incorrect assignments due to translation (such as "*idiot*" associated with the category *plant*). We also include an additional category indicating lemmas that fall within the sphere of sexuality (*sexual words*) which are distinguished from words strictly related to *male genitalia* or *female genitalia*. The resulting lexicon includes 341 of the 501 entries available in the original published resource and a total of 18 lexical categories. The full set of categories, along with their definitions and example lemmas are included in Appendix A

3.2. Datasets and Harmonization

The selection of source datasets is guided by the following criteria: (i) availability of English-language data, (ii) focus on abusive or related phenomena, and (iii) provision of disaggregated annotations, including annotator-level sociodemographic information.

Table 1 provides an overview of the four datasets included in HurtLens. Below we detail what annotators were asked to label in each dataset and how we mapped the original labels onto a binary offensive/not offensive distinction.

- Attitudes (Sap et al., 2022): Annotators were asked *"How toxic/hateful/disrespectful or offensive does this post seem to you?"*, *"How much could this post be seen as toxic/hateful/disrespectful or offensive to anyone?"* and *"In your opinion, how racist is this post?"* on a 1-5 scale.
- Voiced (Weerasooriya et al., 2023): Annotators were generally asked if the text was offensive to them, using a four-point scale (*Not at all Offensive, Slightly Offensive, Moderately Offensive, Very Offensive*).
- SBIC (Sap et al., 2020): Annotators were asked *whether a post could be considered "offensive to anyone"* with three possible categorical answers (*yes, maybe, no*).
- POPQUORN (Pei and Jurgens, 2023): Annotators were asked *"Consider you read the above comment on Reddit, how offensive do you think it is?"*. on a 1-5 Likert scale, where 1 means *"Not offensive at all"* and 5 means *"Very offensive"*.

Given that each dataset adopts its own conceptualization of offensive speech, we perform a harmonization process over both the offensive speech labels and the annotators' demographic variables.

For the offensive speech labels, we reduce the label space to a binary setting, mapping all instances to either the offensive or non-offensive class. Specifically, we adopt a conservative harmonization in which all labels indicating any degree of offensiveness are grouped into the offensive class [3]. This strategy was chosen to retain all possible ways of using a lemma offensively, regardless of the degree, to capture a broader range of potential harmfulness. We further discuss the implications of this choice in Section 7.

For the sociodemographic variables, we normalize heterogeneous annotations into a unified schema. Gender is mapped to four categories (*man, woman, nonBinary, transman*), while for race we consider seven groups (*asian, black, hispanic, white, native, arab, other*). Age is discretized into standard intervals (*18–24, 25–34, 35–44, 45–54, 55–64, 65 or older*). Political ideology is reduced to three categories (*left, right, other*), and education levels are grouped into five (*less_than_high_school, high_school_diploma, bachelors_degree, graduate_degree, other*).

[3] https://github.com/SDavenia/hurt_persp/blob/main/utils/dataset_lexicon_processing.py

Dataset	Platform	#Annotations	#Annotators	#Texts	#Off. Texts	Avg/Annotator	Avg/Text
Attitudes	Twitter	3454	184	627	586(93.0%)	18.77	5.51
Voiced	Reddit	44676	726	2338	2327(∼ 100%)	61.54	19.11
SBIC	Tw./Red./Gab/Storm.	144649	304	45223	30863(68.0%)	377.71	2.54
POPQUORN	Reddit	13036	262	1500	1338(89.0%)	49.76	8.69

Table 1: Summary of datasets with annotation statistics containing unique number of annotators, unique total number of annotations, number of annotators, number of texts and number of offensive texts. We report the number of offensive texts by counting instances where at least one annotator flagged it as offensive.

This harmonization enables consistent cross-dataset comparisons while preserving the core demographic distinctions captured in the original annotations.[4]

3.3. HurtLens

For each HurtLex lemma, we construct HurtLens by retrieving from the selected datasets instances in which the lemma appears, encompassing both hurtful and non-hurtful usages. We use `spaCy` to perform lemmatization and part-of-speech tagging on the texts, and then we match the extracted lemma–POS pairs with the corresponding HurtLex entries.

A visual workflow of HurtLens is reported in Figure 1, while Table 2 reports its main statistics, where triplets correspond to the extracted text-lemma-annotation units.

Dataset	#Triplets	#Texts	#Offensive	#Annotators
Attitudes	3052	364(58.0%)	356(97.8%)	148(80%)
Voiced	41468	1332(57%)	1331 (∼ 100%)	726(100%)
SBIC	59429	18066(40%)	14331(79.3%)	235(77%)
POPQUORN	6665	569(38%)	546(96.0%)	262(100%)
HurtLens	110614	20331(40.9%)	16564(81.5%)	1371(92.9%)

Table 2: Summary of HurtLens statistics, triplets are the extracted text-lemma-annotations. We report also the number of texts extracted, the number and percentage of texts originally annotated as offensive, the number of unique annotators, and the percentage of annotators retained from the original datasets after retrieval.

4. Methodology

We model the offensiveness label, which is our dependent variable y as a binary variable using a generalized linear mixed model, using logistic regression with a logit link function.

[4] https://github.com/SDavenia/hurt_persp/blob/main/data/sociodemographic_mappings.json

4.1. Preprocessing for Modelling

For this analysis, we only consider the sociodemographic variables which are available in all datasets considered, namely *race*, *age* and *gender*. To ensure reliable estimates of the model parameters, we excluded all sociodemographic levels for which any combination with other variables (corresponding to the interaction effects) has less than 30 observations. A similar filtering is conducted on the HurtLex categories. We filter these cases for the purposes of the present analysis; however, the released resource retains them for other downstream uses.

Moreover, we remove all instances where some sociodemographic is set to *other*, as it is deemed not informative. After this filtering step we are left with the following levels for each variable: age (*18-24, 25-34, 35-44, 45-54, 55-64*), therefore excluding *65 or older*; race (*white, black, asian*), excluding *hispanic, native, arab*; gender (*man, woman*), excluding *non-binary, transman*, while for the HurtLex categories we exclude *is_or* (plants), with a total of 17 types of offensive language.

After this filtering step, the dataset upon which we build our models consists of 70062 text-annotator instances, from 1247 unique annotators on 19945 unique texts.

For modeling purposes, both sociodemographic variables and lexical category tags correspond to the fixed effects under investigation for this analysis, while a by-annotator and by-text intercept for *annotator_id* and *text_id* are included as random effects to account for text and annotator variability. Additionally, for the sociodemographic variables we set the reference level to the most common traits, namely *white, man, 25-34*.

4.2. Models Definition

The **null model** (**N**) does not include any fixed effexts and only considers a by-annotator and by-text random intercept. In R notation:

$$y \sim 1 + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Sociodemographic-only Models These models only consider the sociodemographic variables as fixed effects.

For the first model we consider these variables as independent, non-intersecting predictors. We denote this model as the **sociodemographic model** (**S**), in R notation:

$$y \sim \text{race} + \text{age} + \text{gender} + (1|\text{rater\_id}) + (1|\text{text\_id})$$

For the second model, we focus on the interaction between *race* and the other two sociodemographics, grounded in previous literature on intersectionality which showed that it is a common predictor to interact with other variables (Homan et al., 2024). We denote this model as the **sociodemographic race-intersectional model** (**SRi**), in R notation:

$$y \sim \text{race} * (\text{age} + \text{gender}) + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Tags Model We consider a model using each binary variable denoting the presence of lemmas from certain HurtLex categories as independent fixed-effects. We denote this model as the **tag model** (**T**), in R notation:

$$y \sim \text{is\_ps} + ... + \text{is\_re} + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Sociodemographics-Tags Interaction Models For this set of models, we consider both sociodemographic variables and whether the text contains lemmas from certain HurtLex categories.

We first include a model where the race-intersectional model is enriched with the various tags. This model is denoted as **race-intersectional + tags model** (**SRi-T**), in R notation:

$$\begin{aligned} y \sim\ & \text{race} * (\text{age} + \text{gender}) + \\ & + \text{is\_ps} + \cdots + \text{is\_re} \\ & + (1 \mid \text{rater\_id}) + (1 \mid \text{text\_id}) \end{aligned}$$

Finally, we define a model that allows us to investigate how different sociodemographic traits interact with texts containing different types of offensive language, identified via HurtLex categories. Fitting an interaction term between each category tag and sociodemographic variable would lead to an overly-complex model. As such, we conduct an exploratory analysis to identify which interaction terms are of interest.

The methodology for this exploration is described in Section 4.3 with results in Section 5.1, leading to the inclusion of 4 interaction terms between *age* and category tags and 3 for *age*. The final model is denoted as **race-intersectional + tag-sociodemographic model** (**SRi-TS**), in R notation:

$$\begin{aligned} y \sim\ & \text{race} * (\text{age} + \text{gender}) + \\ & + \text{is\_ps} + \cdots + \text{is\_re} \\ & + (\text{is\_asm} + \text{is\_ddp} + \text{is\_ps} + \text{is\_asf}) * \text{race} \\ & + (\text{is\_ddf} + \text{is\_pa} + \text{is\_is}) * \text{age} \\ & + (1 \mid \text{rater\_id}) + (1 \mid \text{text\_id}) \end{aligned}$$

4.3. Exploratory Analysis Methodology

To select meaningful interaction terms for the SRi-TS model, we conduct an exploratory analysis to identify which texts, grouped by the presence of lemmas from specific HurtLex categories, cause different demographic levels to diverge most in their offensiveness ratings.

For every pair of levels of a sociodemographic variable (A, B) and category c, we identify the textual instances containing at least one lemma belonging to category c that were rated by at least one annotator from each group, and obtain a within-group majority label (excluding ties). We compute the divergence between groups A, B on category c as:

$$w^c_{AB} = (n^c_{A>B} - n^c_{B>A})/n_{AB}$$

Where $n^c_{A>B}$ indicates the number of instances annotated as offensive by sociodemographic level A but not B and $n^c_{B>A}$ the opposite, over a total of n^c_{AB} instances with lemmas from category c that were rated by both. This coefficient $w_{AB} \in [-1, 1]$ serves as a sensitivity index: values near 0 indicate consensus, while values toward the extremes indicate that one group consistently perceives those texts as more offensive than the other. To avoid drawing conclusions from a few observations, we exclude pairs with fewer than 20 comparisons, and include interaction terms between the sociodemographic variable and c if there is at least a coefficient $w^c_{AB} > 0.25$. This procedure is exploratory and heuristic rather than a formal model-selection method, and its limitations are discussed in Section 7.

5. Results

5.1. Exploratory Analysis

For our exploratory analysis, we compute w^c_{AB} for every pair of values A, B for a specific sociodemographic variable and category c. We visualize only comparisons for pairs and tags where there is at least a coefficient larger in absolute value than 0.20. Cells highlighted in green indicate that annotators from the first group on the x-axis annotated more instances as offensive, while those in pink indicate the opposite.

Figure 2: w_{AB} values for levels of *race*, only showing pairs and tags where at least one entry is greater or equal than 0.20. A positive value (in green) indicates that the first entry in the comparison annotated more as offensive, while a negative one indicates the opposite.

Figure 2 shows the results for *race*. We observe that across the observed categories *white* annotators tend to label more instances as offensive than both *black* and *native* annotators, and the same behaviour is observed for *asian* annotators compared to *black* ones.

From this exploration we decided to include an interaction term between *race* and *asf, ddp, asm, ps* (corresponding to female genitalia, cognitive disability, male genitalia and negative stereotypes/ethnic slurs). An interaction term with *pr* (words related to prostitution) is not included as it appears above our threshold only in a comparison with *race=hispanic*, which is not included in the model as stated before.

Figure 3 shows the results of the exploratory analysis for *age*. We observe that in general across the observed categories, young annotators tend to label more instances as offensive compared to older ones, with only some exceptions.

Similarly to what we did for *race*, we decided to include interaction terms between *age* and *ddf, is, pa* (corresponding to physical disability, social/economic disadvantage and professions/occupations). An interaction term with *ddp* (corresponding to cognitive disability) is not included since level *65 or older* is excluded from the model.

Concerning *gender*, across all combinations, no w_{AB} is above our pre-specified threshold and no interaction term is included in the model.

5.2. Model Comparison

We report number of degrees of freedom, marginal and conditional R^2, Akaike (AIC) and Bayesian (BIC) Information Criteria across all models in Table 3. Degrees of freedom reflect model complexity in terms of the number of estimated parameters. Marginal R^2 represents the proportion of variance explained by fixed effects alone, while conditional R^2 captures the variance explained by both fixed and random effects. AIC and BIC are information criteria used for model comparison, balancing goodness of fit with model complexity; lower values indicate a better trade-off between fit and parsimony.

Model	df	M-R^2(↑)	C-R^2(↑)	AIC(↓)	BIC(↓)
N	3	—	81.1	62661	62689
S	10	3.2	81.6	62592	62683
SRi	20	5.3	81.7	62573	62756
T	19	4.2	81.6	61189	**61363**
SRi-T	36	9.3	**82.2**	61101	61431
SRi-TS	56	**9.4**	**82.2**	**61087**	61600

Table 3: Model comparison statistics, reporting number of degrees of freedom (df), marginal and conditional R^2 (M-R^2, C-R^2) as percentage of total variability, AIC and BIC.

The results from the null model reveal that 81.1% of variability can be accounted for by the random effects, indicating that a large part of variance is explained by text-specific and annotator-specific characteristics.

The results from the sociodemographic models indicate that sociodemographic variables treated independently (S) account for 3.2% of the total variability. However, when modeled with an intersectional approach (SRi) they account for 5.3% of total variability, **emphasizing the importance of considering interaction effects with an intersectional approach**. These findings agree with Hu and Collier (2024), who observed that sociodemographic variables accounted for 4.5% and 2.9% on AnnwithAttitudes and POPQUORN respectively. Similarly, in line with previous work on intersectionality by (Homan et al., 2024), when considering the intersection of *race* with the other sociodemographic variables the explained variability increases.

The results from the tag model (T) indicate that the presence of specific HurtLex lemma category tags accounts for 4.2% of total variability when treated independently. This effect is similar in magnitude to that of the sociodemographic variables.

When considering the models leveraging both tags and sociodemographics information (SRi-T,

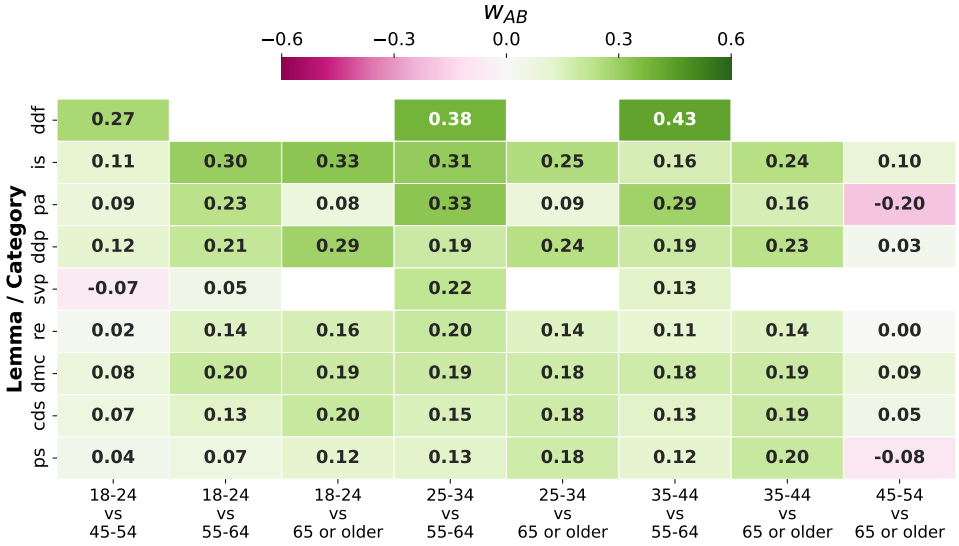

Figure 3: w_{AB} values for levels of *age*, only showing pairs and tags where at least one entry is greater or equal than 0.20. A positive value (in green) indicates that the first entry in the comparison annotated more as offensive, while a negative one indicates the opposite.

SRi-TS), the variability explained by those factors jumps to 9.3% for the model not considering sociodemographic-tags interactions and 9.4% for the other. This increase compared to both tags-only and sociodemographic-only models indicates that the sociodemographics and presence of specific lemma category tags model different aspects of the variability.

While the interaction model (SRi-TS) only results in a small increase in marginal R^2, we choose to utilize this model as it allows us to investigate the interaction between sociodemographic groups and the presence of lemmas from HurtLex belonging to the identified categories. Moreover, while BIC favours simpler models due to its penalty for larger parameters, the decrease in AIC compared to all other models justifies the inclusion of the interaction terms of the SRi-TS model.

5.3. Effects and Interaction of Sociodemographics and Tags

For the rest of the analysis, the predicted probabilities of offensive ratings are obtained with the SRi-TS model using the Average Marginal Effect (AME) method via the `ggeffects` package `predict_response` function with `margin="average"`, ensuring the predicted probabilities are averaged over the distribution of all observations (Lüdecke, 2018). Additionally, we include the 95% confidence interval error bars for these predicted probabilities.

Effect of Sociodemographics To answer **RQ1**, we investigate how the effect of *age* and *gender* varies across different levels of *race*. This allows us to visualize the intersectional component of the model, allowing for the influence of *age* and *gender* to vary depending on the annotator's *race*.

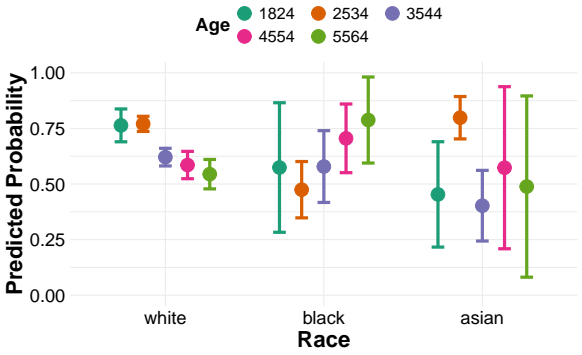

Figure 4: Predicted probability of offensiveness rating (AME) for *race* and *age* interaction.

Figure 4 shows the effect of *age* across different levels of *race*. We note that while for *white* annotators older age groups have lower predicted probability of rating a text as offensive than younger groups, an opposite trend is observed on average when focusing on *black* annotators, with only little overlap between the error bars for $25-34$ and $55-64$ age groups. Within *asian* annotators, we observe that the $25-34$ group is identified as being the one most likely to rate content as offensive. However, for the other groups, particularly the older ones, the limited number of observations leading to large confidence intervals does not allow us to derive strong conclusions.

Similarly, Figure 5 shows the effect of *gender*

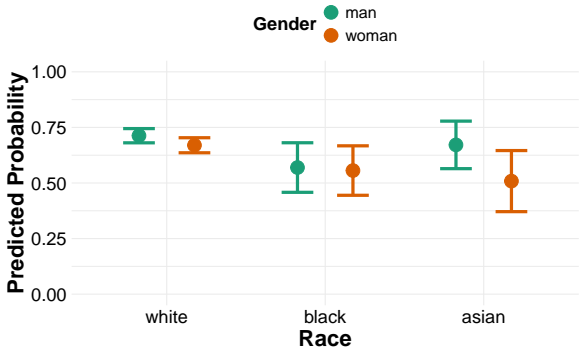

Figure 5: Predicted probability of offensiveness rating (AME) for *race* and *gender* interaction.

across different levels of *race*. We observe that across both *white* and *asian* annotators, women appear to be less likely to rate texts as offensive, while this behaviour is not observed in *black* annotators.

Effect of Tags To answer **RQ2**, we investigate how the presence of lemmas belonging to certain HurtLex Categories impacts the predicted probability of offensiveness of those texts. Figure 6

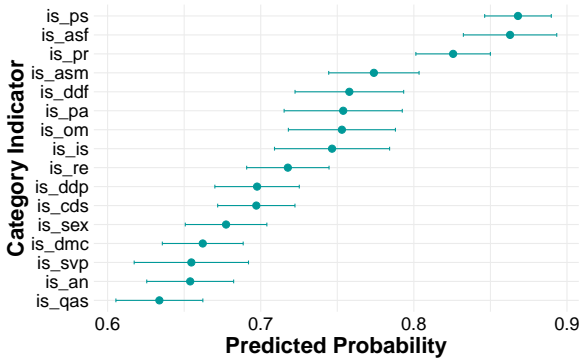

Figure 6: Predicted probability of offensiveness rating (AME) for texts containing lemmas from different HurtLex category tags.

shows how the predicted probability of a text being offensive changes based on which HurtLex tag the text contains. We observe that texts containing words related to stereotypes, slurs (*is_ps*), female genitalia (*is_asf*) and prostitution (*is_pr*) are much more likely than the others to be annotated as offensive, with the CI lower bound above 0.80. On the contrary, texts containing potentially negative words (*is_qas*), animals (*is_an*), seven deadly sins (*is_svp*) and moral/behavioral defects (*is_dmc*) are less likely than the others to be annotated as offensive, with the upper bound of the CI falling below 0.7.

Effect of Sociodemographics and Tags Interactions Finally, to answer **RQ3** we investigate how different sociodemographic groups vary in their offensiveness ratings on texts containing words belonging to different HurtLex categories. For our SRi-TS model, described above, we included 3 interaction effects with *age* and 4 with *race*. Here we report only those where we observed at least a significant interaction effect compared to the base group (i.e. $25-34$ for *age* and *white* for *race*).

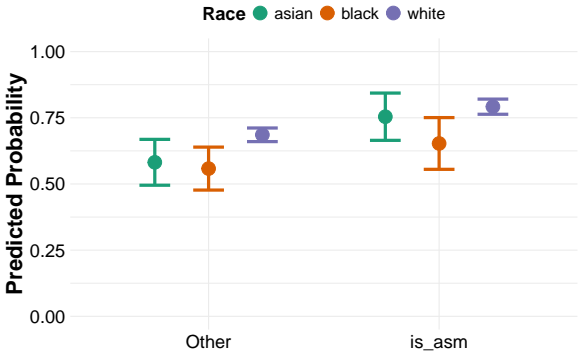

Figure 7: Predicted probability of offensiveness rating (AME) for texts containing words related to male genitalia (*is_asm*) for varying levels of *race*.

Figure 7 shows how annotators with different *race* change their ratings comparing texts without male-genitalia related words to those containing them. We observe that while across all groups the presence of *is_asm* (male genitalia) words increases the predicted probability of offensiveness ratings, this increase is more pronounced for *asian* annotators. In particular, *asian* annotators are more similar to *black* annotators when those terms do not appear, and less likely than *white* annotators to identify texts as offensive, but their offensiveness ratings are more similar to *white* annotators for texts containing *asm* related lemmas. This indicates that *asian* annotators appear to be particularly sensitive to lemmas belonging to this category. This agrees with our exploratory findings where we did not identify any difference between *white* and *asian* annotators on these terms.

Similarly, Figure 8 shows how annotators with different *race* change their ratings comparing texts with cognitive-disability related words to those containing them. We observe different effects across the different *race* groups, where *asian* and *white* annotators become slightly more likely to identify texts as offensive when they contain *ddp* (cognitive disability) lemmas, while for *black* annotators the opposite behaviour is observed. This complements our exploratory findings, where we observed that both *asian* and *white* annotators were more likely than *black* annotators to annotate as offensive texts containing cognitive-disability related

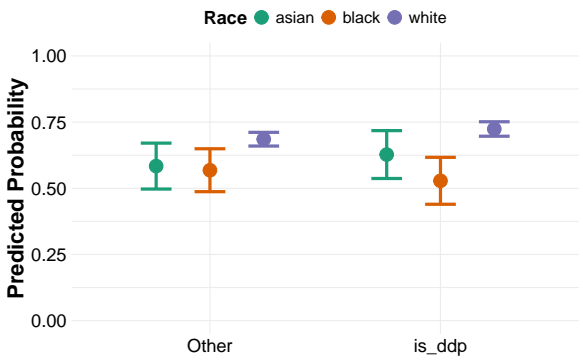

Figure 8: Predicted probability of offensiveness rating (AME) for texts containing words related to cognitive disability (*is_ddp*) for varying levels of *race*.

words.

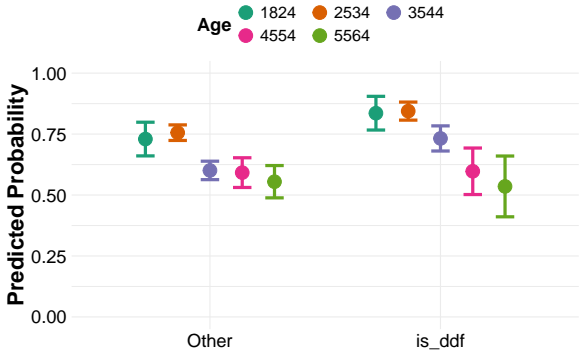

Figure 9: Predicted probability of offensiveness rating (AME) for texts containing words related to physical disability (*is_ddf*) for varying levels of *age*.

Finally, Figure 9 shows how annotators with different *age* change their ratings comparing texts with physical-disability related words to those containing them. We observe different effects across various *age* groups. In particular, while for younger age groups (up to $35-44$ group), the annotators are more likely to rate texts containing *ddf* (physical disability) lemmas as offensive, this change is not observed in the older demographics. This points to younger generations being more sensitive to the topic of physical disability.

6. Conclusions

In this work, we introduce HurtLens by enriching existing disaggregated datasets with hurtful lexicon-based information. We demonstrate that offensiveness perception is shaped by both the sociodemographic characteristics of annotators, the lexical categories of hurtful words present in texts and the interaction between them. Through the proposed mixed-effects modelling approach (SRi-TS) we found that intersectional modelling increased explained variance (marginal R^2) compared to models including independent sociodemographic predictors. In particular, *race* and *age* groups exhibited different rating shifts depending on the presence of specific lexical categories, uncovering **novel insights into group-specific sensitivities to distinct types of hurtful language.**

These findings underscore the importance of adopting perspectivist approaches to offensive language detection.

Moreover, our methodology is generalizable to other pragmatic phenomena, provided a prior structured knowledge, e.g., a lexicon capturing the target phenomenon, and suitable disaggregated source datasets.

As part of future work, we intend to further investigate not only the categories but also the individual lemmas in HurtLex, examining how the perception of specific lemmas varies according to sociodemographic variables and context of use. By shifting the analysis to the lemma level, it will also be possible to incorporate additional lexical resources in order to expand the lexical coverage.

Furthermore, we aim to leverage perspective-specific examples from HurtLens to tackle over-moderation issues in Large Language Models on offensive speech detection. Current systems often struggle to identify hurtless usage of certain lemmas, leading to many False Positives (Draetta et al.). By utilizing a perspectivist corpus which also includes non-offensive usage of certain words and demographic-specific interpretations could help models disambiguate between different uses of these words.

Finally, we intend to adopt community-led and participatory approaches to validate and refine the resource with input from the demographic groups represented in our analyses.

7. Ethics Statement and Limitations

Ethics Statement The primary goal of HurtLens is to provide a resource for combating online hate by enabling more nuanced models that can better account for the multifaceted nature of offensiveness perception across different demographic groups, ultimately supporting more equitable content moderation systems. We acknowledge that resources documenting hurtful language carry inherent risks of misuse. HurtLens is intentionally designed to expose the diversity of perspectives on offensiveness rather than to provide a definitive catalog of harmful terms. The resource should not be used to target or harass specific demographic groups, nor to train models that disproportionately silence marginalized voices. Our perspectivist approach is motivated by the goal of reducing bias

in content moderation by making systems aware of how different communities perceive harm differently, thereby avoiding both over-moderation of minority perspectives and under-moderation of actual harmful content. We strongly discourage any application of this resource that would amplify harm or reinforce existing power imbalances in online spaces.

Limitations We identify limitations of our work. First, it is restricted to the English language, which may limit the generalization of the findings to other linguistic and cultural contexts. Similarly, our methodology builds upon the HurtLex lexicon, and therefore inherits its coverage limitations despite our revision efforts. Additionally, the resource focuses on explicit lexical realizations of hurtful language and does not account for implicit expressions of hate or stereotypes, which often require deeper contextual and pragmatic interpretation. Furthermore, our use of spaCy for lemmatization may fail to resolve non-standard slang or intentional misspellings.

From a modeling perspective, our selection of interaction terms between sociodemographic levels and HurtLex categories was guided by an exploratory analysis rather than an exhaustive search using more appropriate model selection criteria. While this approach allowed us to identify salient interaction terms, it is not exhaustive. Moreover, we did not consider intersectional groups interactions with lexical categories, which could capture additional patterns. As a methodological note, employing Bayesian approaches could provide more robust and interpretable estimates, but given the dataset size we were limited by computational requirements for this exploratory analysis.

Another limitation arises as we reduced heterogeneous annotation schemes to a binary offensive vs. non-offensive label which may hide important differences in the original scales. In particular, our conservative choice to treat any degree of offensiveness as offensive merges mild, ambiguous, and severe cases into a single case. While this supports a broad view of harmful usage, different thresholding choices (e.g., excluding mid-scale values) could lead to different distributions and effects.

Finally, combining multiple source datasets introduces potential confounds, as they differ in annotation guidelines, platform, and annotator composition. As a result, some observed demographic or lexical effects may partly reflect dataset-specific artifacts rather than general patterns. While our unified analysis aims at capturing broader trends, controlling for dataset effects (e.g., via per-dataset models or dataset indicators) could help disentangle these factors, and we leave this for future work.

Acknowledgements

This work was funded by the partnership with Amazon Science "Multilingual personalization through perspective-aware Language Modeling" and the project NEIKEA (Bando CSP TRAPEZIO - Linea 1 - Paving the way to research excellence and talent attraction).

8. Bibliographical References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Inf.*, 13(6):273.

Iván Árcos and Jaime Pérez. 2023. Detecting hurtful humour on twitter using fine-tuned transformers and 1d convolutional neural networks. In *IberLEF@SEPLN*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile. 2021. *It's the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks*, page 441–453. Springer International Publishing.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 52–57, Turin, Italy. CEUR Workshop Proceedings.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023a. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023b. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena of the Italian Chamber of Deputies, which issued a Final Report in 2017.

Lia Draetta, Soda Marem Lo, Samuele D'Avenia, Valerio Basile, and Rossana Damiano. Testing llms' sensitivity to sociodemographics in offensive speech detection.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Paula Fortuna, Juan Soler Company, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. pages 6786–6794.

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.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Luca Giordano and Maria Pia Di Buono. 2023. Assessing Italian news reliability in the health domain through text analysis of headlines. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 538–548, Vienna, Austria. NOVA CLUNL, Portugal.

Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2024. Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia. ELRA and ICCL.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.*, 71:431–478.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Daniel Lüdecke. 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26):772.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Anais Ollagnier. 2024. CyberAgressionAdo-v2: Leveraging pragmatic-level information to decipher online hate in French multiparty chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4287–4298, Torino, Italia. ELRA and ICCL.

Petya Osenova. 2024. On a hurtlex resource for Bulgarian. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 214–219, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory

for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

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.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. 2020. Multi-word expressions for abusive speech detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online. Association for Computational Linguistics.

Alice Tontodimamma, Lara Fontanella, Stefano Anzani, and Valerio Basile. 2023. An italian lexical resource for incivility detection in online discourses. *Quality & Quantity: International Journal of Methodology*, 57(4):3019–3037.

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.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate - A systematic review of hateful communication datasets. *CoRR*, abs/2405.08562.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

9. Language Resource References

Bassignana, Elisa and Basile, Valerio and Patti, Viviana. 2018. *HurtLex*. PID https://github.com/valeriobasile/hurtlex.

Pei, Jiaxin and Jurgens, David. 2023. *POPQUORN*. PID https://github.com/Jiaxin-Pei/potato-prolific-dataset.

Sap, Maarten and Gabriel, Saadia and Qin, Lianhui and Jurafsky, Dan and Smith, Noah A. and Choi, Yejin. 2020. *SBIC*. PID https://huggingface.co/datasets/allenai/social_bias_frames.

Sap, Maarten and Swayamdipta, Swabha and Vianna, Laura and Zhou, Xuhui and Choi, Yejin and Smith, Noah A. 2022. *Attitudes*.

Weerasooriya, Tharindu and Dutta, Sujan and Ranasinghe, Tharindu and Zampieri, Marcos and Homan, Christopher and KhudaBukhsh, Ashiqur. 2023. *Voiced*. PID https://huggingface.co/datasets/Lab-PL/voiced.

A. Additional Details

A.1. HurtLex categories

Column Tag	Definition	Example
ps	negative stereotypes dearethnic slurs	jewish
pa	professions and occupations	cop
ddf	physical disabilities and diversity	disabled
ddp	cognitive disabilities and diversity	dumbass
dmc	moral and behavioral defects	liar
is	words related to social and economic disadvantage	poor
or	plants	melon
an	animals	snake
asm	male genitalia	dick
asf	female genitalia	pussy
pr	words related to prostitution	slut
om	words related to homosexuality	twink
qas	with potential negative connotations	camp
cds	derogatory words	baby
re	felonies and words related to crime and immoral behavior	abuse
svp	words related to the seven deadly sins of the Christian tradition	rage
sex	words related to sexual acts	fuck

Table 4: Hurtlex categories with their definitions and example lemmas.