# Not Just Who *or* What: Modeling the Interaction of Linguistic and Annotator Variation in Hateful Word Interpretation

**Sanne Hoeken**[1], **Özge Alaçam**[1], **Dong Nguyen**[2], **Massimo Poesio**[2,3], **Sina Zarrieß**[1]
[1]Bielefeld University, Germany [2]Utrecht University, the Netherlands
[3]Queen Mary University of London, United Kingdom
{sanne.hoeken, oezge.alacam, sina.zarriess}@uni-bielefeld.de
{d.p.nguyen, m.poesio}@uu.nl

## Abstract

Interpreting whether a word is hateful in context is inherently subjective. While growing research in NLP recognizes the importance of annotation variation and moves beyond treating it as noise, most work focuses primarily on annotator-related factors, often overlooking the role of linguistic context and its interaction with individual interpretation. In this paper, we investigate the factors driving variation in hateful word meaning interpretation by extending the HateWiC dataset with linguistic and annotator-level features. Our empirical analysis shows that variation in annotations is not solely a function of *who* is interpreting or *what* is being interpreted, but of the interaction between the two. We evaluate how well models replicate the patterns of human variation. We find that incorporating annotator information can improve alignment with human disagreement but still underestimates it. Our findings further demonstrate that capturing interpretation variation requires modeling the interplay between annotators and linguistic content and that neither surface-level agreement nor predictive accuracy alone is sufficient for truly reflecting human variation.[1]

**Content warning!** Some examples in this paper contain language that may be offensive, for illustrative purposes; we recognize their potential harm.

## 1 Introduction

Words play a central role in hate speech by encoding derogatory meanings. The meaning of such words is rarely fixed but highly dependent on context and interpretation which poses a significant challenge for both theoretical understanding and computational modeling of hate speech (Sayeed, 2013). Despite growing interest in hate speech detection, there has been little systematic investigation into the semantic and pragmatic mechanisms that underlie how hateful word meanings are interpreted.

Recent work by Hoeken et al. (2024) introduced the HateWiC dataset, which identified substantial variation and disagreement in judgments about whether a word is hateful in context. Models tend to underperform on those cases where annotators disagree. Although incorporating annotator demographic information shows modest improvements in model performance, the underlying sources driving these variations remain poorly understood. This aligns with a broader trend in NLP research, that moves away from aggregated judgments to explicitly modeling inter-annotator variation (Uma et al., 2021; Basile et al., 2021).

Yet, the focus in NLP research on label variation in subjective tasks has largely remained on *who* is interpreting (Kocoń et al., 2021; Orlikowski et al., 2023), with far less attention given to *what* is being interpreted. While linguistic content has always been the basis for classification, recent subjectivity-focused approaches tend to sideline the role of the content itself. Only a few studies acknowledge the role of linguistic ambiguity in subjective labeling (Sandri et al., 2023; Jiang and Marneffe, 2022).

Table 1 illustrates how subjective variation can emerge from both linguistic and annotator features with examples from HateWiC. Variation in perceived hatefulness of the word *napoleon* in the first example likely arises from ambiguity between senses (food vs. person) with limited context. Whereas in the second example annotator differences likely contributed to disagreement, as the annotators seem to have different tendencies to label content as hateful (based on their label ratios). Lastly, the *shrink* example shows that the same annotator's tendency can shift depending on the

[1]Code and supplementary materials for this study are available at https://github.com/SanneHoeken/HateWiCVariation.



*Proceedings of the 16th International Conference on Computational Semantics*, pages 63–77
September 22-23, 2025, 

| Word in Context | Term | Sense Definition | Sense Domain | Sense Person Aspect | Context Length | Ann. Id | Gender | Hateful Ratio | Label |
|---|---|---|---|---|---|---|---|---|---|
| Miss Manvers thrust aside a garnished plate and attacked her *napoleon*. | napoleon | Another name for a millefeuille pastry. | Food | NotPerson | 11 | 36<br>69 | Female<br>Female | ▲▲▲<br>▲▲ | ✗<br>✓ |
| He is the *napoleon* of crime, Watson. He is the organizer of half that is evil [...] | napoleon | A person having come to dominate an area of activity through illegality. | Person | Personality/ behavior | 41 | 36<br>75 | Female<br>Male | ▲▲▲<br>▲ | ✗<br>✓ |
| My *shrink* said that he was an enabler, bad for me. | shrink | A psychiatrist or psychotherapist. | Person | Profession | 11 | 36<br>4 | Female<br>Female | ▲▲▲<br>▲▲▲ | ✓<br>✗ |

Table 1: Examples from the HateWiC dataset, with augmented linguistic and annotator information, that illustrate how label variation (✗= *hateful*; ✓= *not hateful*) can arise from linguistic ambiguity (e.g. different senses of napoleon) as well as from annotator tendencies (Hatefulness Ratio from *low* (▲) to *high* (▲▲▲)), while also highlighting the interaction of these features with subjective interpretation.

linguistic content they are judging, such as whether the term's referent is defined by profession or behavior (Person Aspect). It is this interaction between linguistic features and subjective tendencies that shapes variation in interpretation.

Within the ongoing search for meaningful predictors of human variation in subjective language interpretation, relatively little attention has been given to the level of word meaning. Moreover, most studies only focus on annotator-related features, neglecting the interplay between semantics, linguistic context, and subjective interpretation that shapes how hateful meanings arise. Additionally, existing modeling efforts typically emphasize overall performance metrics without assessing whether models replicate the *patterns* of human variation. Yet understanding and modeling such patterns is crucial for NLP systems to meaningfully reflect the subjective nature of language interpretation in sensitive tasks like hate speech detection.

Addressing these gaps, we augment the HateWiC dataset with linguistic and annotator-level features (§3) and empirically show that variation in hateful meaning interpretation is driven not just by who the annotator is or what is being annotated, but by their interaction (§4). Building on this analysis, we propose an evaluation framework that assesses whether BERT-based classification models capture this variation (§5). The results (§6) demonstrate that while models incorporating annotator-specific inputs can reproduce superficial variation, they substantially underestimate its magnitude and fail to capture the internal structure of variation found in human annotations.

## 2 Related Work

In what follows, we discuss prior work on hateful word meaning in NLP and subjective variation in Hate Speech Detection (HSD), both of which motivate our study.

### 2.1 Hateful word meaning in NLP

Capturing variation in word meaning has long been a focus in NLP (Pustejovsky, 1991; Schütze, 1998; Haber and Poesio, 2024). Computational approaches to lexical semantics have included tasks such as Word Sense Disambiguation (Loureiro et al., 2021), Word Sense Induction (Eyal et al., 2022) and Lexical Semantic Change Detection (Periti and Montanelli, 2024). Methods predominantly rely on embedding-based techniques using encoder-based language models and often employ contextualized sense similarity metrics (Blevins and Zettlemoyer, 2020; Cassotti et al., 2023). Moreover, the tasks and approaches typically depend on general-purpose resources and corpora that are oriented toward standard language usage. Consequently, they tend to focus on denotative rather than domain-specific or connotative meaning (Potts, 2007) (e.g. capturing denotative shifts as with a word like *plane* changing from primarily a geometric concept to also denoting an aircraft, in contrast to connotative shifts, such as *spinster* becoming more negatively charged over time).

In contrast, some work has addressed connotative meaning in the context of hate speech by examining lexical features used in sequence-level detection (Koufakou et al., 2020; Zampieri et al., 2022). Other studies have explored the disambiguation and detection of such terms, including subtle forms like dog whistles (Kruk et al., 2024; Mendelsohn et al., 2023). Prior research has also examined more clear-cut cases, such as swear words (Pamungkas et al., 2022) and slurs (Hoeken et al., 2023), which are often argued to be more stable across contexts (Frigerio and Tenchini, 2019). Additional work has addressed more ambiguous pejorative terms (Dinu et al., 2021). However, much of this research adopts a (binary) classification perspective, with limited attention to intra-word variation, i.e. how the connotative meaning of a term shifts across con-

texts or individuals. Recently, Hoeken et al. (2024) addressed this issue with the introduction of the HateWiC dataset. Their findings highlight the substantial variation in how hateful word meanings are perceived, but the question about what underlies this variation remains.

### 2.2 Subjective variation in HSD

Annotator disagreement is increasingly recognized as a signal of subjective variation rather than mere labeling noise (Larimore et al., 2021; Plank, 2022; Fleisig et al., 2024). This shift is especially pertinent in HSD, where personal differences strongly influence interpretive judgments. Several studies have highlighted the role of annotator identity in shaping perceived offensiveness. While some highlight the relevance of sociodemographic variables like gender and age (Kocoń et al., 2021; Sang and Stanton, 2022), recent findings suggest that such variables often act as noisy proxies and are poor predictors for interpretation variation (Alipour et al., 2024; Orlikowski et al., 2023). Several studies consider other annotator factors like ideology (Sap et al., 2022) or moral values (Mostafazadeh Davani et al., 2024), yet all consider annotator information as the primary source of variation.

Recent modeling approaches have incorporated annotator-specific information in various ways. These include demographic-based embeddings (Fleisig et al., 2023), embeddings based on annotator ids or label histories (Deng et al., 2023; Mokhberian et al., 2024), and label distribution learning (Weerasooriya et al., 2023). Other recent personalization techniques involve multimodal signals like gaze (Alacam et al., 2024), or fine-tuning LLMs with annotator-specific prompts (Orlikowski et al., 2025). Despite these advances, most efforts emphasize improvements in predictive performance, often evaluated via accuracy metrics. An exception is Anand et al. (2024), who propose aligning model confidence with annotator agreement as a step toward more human-aligned predictions.

Our work contributes to this line of research by explicitly modeling individual variation in hateful word interpretation, and evaluating models by how well they capture the *structure* of this variation across linguistic and annotator-related dimensions.

## 3 Data & Features

To analyze variation in the interpretation of potentially hateful words, we use the HateWiC dataset (Hoeken et al., 2024), which provides contextualized word usages annotated for perceived hatefulness. We further enrich this dataset with additional linguistic and annotator-related features to facilitate a comprehensive empirical analysis of variation.

### 3.1 The HateWiC dataset

The HateWiC dataset comprises approx. 4,000 word-in-context (WiC) instances, each annotated independently by three annotators (N $\approx$ 12k total annotations). Annotation was distributed across 48 annotators, with each annotating 250 instances. Each instance consists of a target term embedded in a sentence and linked to a Wiktionary definition that corresponds to its contextual meaning (totaling 1,888 unique definitions). This setup thus provides sense-level information. The terms included have at least one sense referring to people and considered offensive based on Wiktionary data. Annotators were asked to indicate whether the meaning of the target term in the specific sentence was hateful or not, and could also indicate undecided. The dataset is balanced across the two main classes.

To measure variation, we use a binary variable indicating whether an individual annotator's label matches the majority label for that instance (agree) or not (disagree). We adopt this annotation-level operationalization because it allows us to associate both linguistic features of the text and annotator features (which require individual annotations) with variation in interpretation. We further augment the HateWiC data with various supplementary features, described (and highlighted in bold) below.

### 3.2 Linguistic features

We manually annotated the semantic **Domain** of each Wiktionary definition, assigning categories such as Person, Animal, and Food. This is motivated by the idea that ambiguity across these broad semantic domains (e.g. Napoleon as a person versus a dessert) may lead to variation in hateful interpretation. We further annotated the **Person Aspect** emphasized, distinguishing among categories such as Personality/Behavior, Ethnicity/Nationality and Appearance. These dimensions could influence annotators' judgments of hatefulness differently. For example, references to ethnicity may evoke stronger perceptions of offense compared to those focused on behavior or appearance. All annotations were carried out by two linguistic experts, with full dual annotation for validation. More details on category taxonomies and annotation are provided in

Appendix A.

In addition to these semantic annotations, we included the **part of speech (POS)** linked to each sense definition, which was already included as metadata in the HateWiC dataset. We also consider for each word in context the **Context Length**, measured by the number of whitespace-separated tokens, as shorter contexts might provide fewer clues for disambiguation which potentially increases disagreement among annotators.

Finally, we incorporate the **Grammatical Role** of the target word in its context. Grammatical Role was identified using SpaCy's dependency parser and mapped to a coarser set of ten categories such as subject, object and preposition (fully specified in Appendix A). This syntactic information might affect how strongly a term is emphasized and thereby influence variation in perceived hateful intent.

### 3.3 Annotator features

We incorporated annotator-related features by leveraging information already present in the HateWiC dataset. This includes the **Annotator Id**, along with available sociodemographics (**Gender**, **Ethnicity**, and **Age**). We converted absolute age values into age categories (e.g. '20-29'). As an additional feature, we computed each annotator's **Hatefulness Ratio**, defined as the proportion of instances they labeled as hateful across the dataset (see also Appendix A). This metric serves as an approximation of an annotator's tendency to classify content as hateful.

## 4 Empirical Analysis

We begin our empirical analysis by assessing the overall degree of annotator agreement in the HateWiC dataset. We calculate inter-annotator agreement on the original dataset's annotations, with Krippendorff's alpha resulting in 0.452. This value reflects moderate agreement and matches the original HateWiC paper's findings (Hoeken et al., 2024)[2]. Moving beyond surface-level agreement, we statistically test the association between our enriched set of linguistic and annotator features, and the binary outcome of agreement with the majority.

### 4.1 Independent feature associations

For a fair comparison of statistical test outputs, we converted numerical features (Context Length and Hatefulness Ratio) into categorical variables using quantile-based binning (with n_bins = 4). We conducted Chi-squared tests of independence to assess the relationship between each feature and annotation agreement (i.e. agree or disagree with the majority vote). Effect sizes were calculated using Cramer's V to measure the strength of associations.

| Type | Feature | $\chi^2$ | p-value | Cramer's V |
|---|---|---|---|---|
| **linguistic** | Person Aspect | 61.43 | <0.001 | 0.072 |
| | Domain | 31.83 | <0.001 | 0.052 |
| | Context Length | 48.53 | <0.001 | 0.064 |
| | Grammatical Role | 18.99 | 0.040 | 0.040 |
| | POS | 4.06 | 0.669 | 0.018 |
| **annotator** | Annotator Id | 238.11 | <0.001 | 0.141 |
| | Hatefulness Ratio | 37.32 | <0.001 | 0.056 |
| | Ethnicity | 59.39 | 0.000 | 0.071 |
| | Age | 14.53 | 0.006 | 0.035 |
| | Gender | 4.73 | 0.094 | 0.020 |

Table 2: Statistical test results for association of categorical features with annotation agreement (agree or disagree with majority vote)

The results in Table 2 show several statistically significant associations. Among linguistic features, Person Aspect shows the strongest association. Context Length and Domain also have significant effects on the agreement ($p < 0.001$) and Grammatical Role is marginally significant ($p = 0.04$). In the annotator-related features, Annotator Id shows the strongest association. Ethnicity and Hatefulness Ratio are also significant ($p < 0.001$). Age is significant at the 0.01 level. Further details on the computation and results, including contingency tables, are provided in Appendices B and D.

Overall, the analysis indicates that both linguistic properties of the input and demographic/behavioral characteristics of annotators influence annotation variation, with the strongest effects observed at the annotator level. While many features have significant effects, the effect sizes are generally small (Cramer's V $< 0.15$), indicating weak to modest associations. This suggests that a large portion of variation in annotation variation remains unexplained by these main effects.

### 4.2 Feature interaction associations

Figure 1 displays both individual and pairwise interaction effects on annotation agreement, again based on Chi-squared tests, this time considering combinations of two features as well. The diagonal represents individual feature effects, while the off-diagonal quadrants correspond to pairwise interactions: the lower-left quadrant shows interactions between linguistic and annotator features, the

[2]The alpha value reported in Hoeken et al. (2024) was obtained without considering the undecided label, a difference that does not appear to substantially affect the outcome.

upper-left linguistic × linguistic interactions, and the lower-right annotator × annotator interactions.

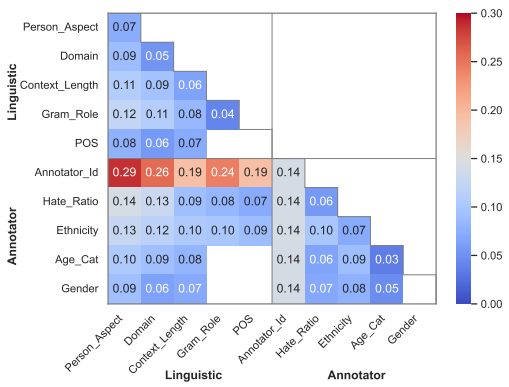


Figure 1: Heatmap of Cramer's V effect sizes showing both individual and pairwise associations of features with annotation agreement. The upper triangle (above the diagonal) as well as non-significant ($p > 0.05$) interaction effects are masked.

Generally, interactions explain more variation in annotation agreement than individual features. Particularly, interactions between annotator and linguistic features are the strongest, with the highest effect size of V = 0.29 for Person Aspect × Ann Id. This pattern of low main effect but high cross-type (linguistic × annotator) interaction supports that annotation variation is more a function of who is interpreting what, rather than just who, or what.

Within type interactions, the higher interaction effects among linguistic features (max V = 0.12 for Person Aspect × Grammatical Role), compared to individual features (max V = 0.07), emphasize that the combined effect of linguistic features matters more for meaning variation, which aligns with linguistic theories of compositionality and context-dependent meaning (Partee et al., 1984).

Adding interactions among annotator features does not increase association strength beyond what is captured by Annotator Id alone. This is logical because Annotator Id essentially encapsulates all annotator-related factors. Ignoring Annotator Id, interactions among other annotator features show modestly stronger effects than individual features, with the ethnicity × Hatefulness Ratio interaction yielding V = 0.10. This implies possible interpretative biases (reflected by tendency to label hate) linked to cultural context. Nonetheless, these effects remain smaller than those involving Annotator Id, thus the results show that individual annotator differences beyond demographics and labeling tendency has stronger influence on the agreement.

### 4.3 A closer look: Person Aspect × Hatefulness Ratio

While statistical tests and interaction analyses provide evidence of feature associations with annotation agreement, inspecting the directions and patterns of these effects allows for a more concrete interpretation. We illustrate this by zooming in on the interaction effect of two features from our analysis. Figure 2 visualizes the interaction between the semantic Person Aspect of the target word and annotators' hateful labeling tendency (Person Aspect × Hatefulness Ratio, with the latter discretized into four intervals. The disagreement probability shows distinct patterns across Person Aspect categories. For example, instances in the Appearance or Social class categories exhibit relatively high disagreement for annotators with a low Hatefulness Ratio and less disagreement with moderate to high ratios. Conversely, the Kinship/social category exhibits the opposite trend. These diverging patterns emphasize that annotator tendencies do not exert uniform effects across linguistic categories. Instead, the influence of individual biases on annotation variation is mediated by the specific semantic characteristics of the content.

## 5 Computational Modeling

In this section, we investigate to what extent computational models with different inputs can capture human variation in annotations. We address this question in the context of the binary classification task that predicts the individual annotations in the HateWiC dataset (12K annotations of words in context, labeled hateful or not hateful based on their meaning in that context). We explicitly model and analyze this variation by conditioning predictions on auxiliary inputs such as annotator identity or demographics. The primary goal is to gain insights into alignment with human interpretation variation rather than optimize benchmark performance.

### 5.1 Model architecture & experiments

We largely follow the approach proposed by Deng et al. (2023), who incorporate annotator embeddings into a BERT model. Their mechanism relies on a predefined annotator id vocabulary. We extend this approach by introducing a modular framework that allows integration of *auxiliary* information, including not only discrete id-based inputs but also free-form text descriptions, alongside standard input text (*primary* input). The model architecture

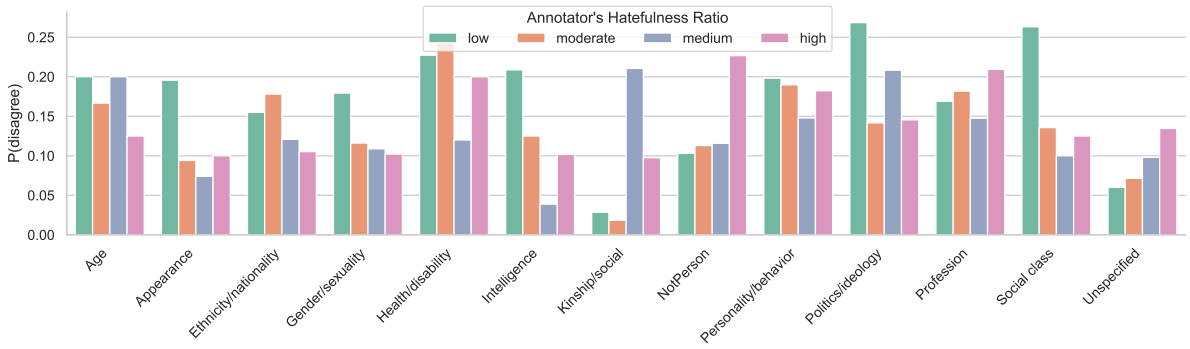


Figure 2: Interaction between the person-related semantic category of the target word (Person Aspect) and annotator's individual tendency to label instances as hateful (Hatefulness Ratio) on the probability of disagreement as the proportion of annotations where individual annotators disagreed with the majority vote.

builds upon a pre-trained encoder for representations of textual inputs. Specifically, we initialize all models with the base version of ModernBERT (Warner et al., 2024) as encoder. Similar to Deng et al. (2023) we adopt a learnable feature-wise weighting mechanism for auxiliary embeddings.

**Primary text embeddings** For each HateWiC instance, the primary input is the sentence containing the target term (**WiC**). Alternatively, following Hoeken et al. (2024), we test replacing this input with the corresponding Wiktionary definition (**WikDef**), or using a concatenation of both. WikDef provides lexical semantic information about the term in a non-contextualized form. Each input type is independently passed through the encoder to obtain a [CLS] representation, which serves as the primary feature embedding.

**Auxiliary annotator embeddings** Following Deng et al. (2023), an embedding layer maps auxiliary ids to dense vectors, which are jointly trained with the rest of the model, yielding *id-based* annotator embeddings (**ann id**). We extend this framework by enabling auxiliary inputs in natural language form, resulting in *text-based* annotator embeddings. These include: (i) annotator ids (**ann id**) expressed as text (e.g. "annotator_12"), (ii) a description of demographic characteristics (**profile descr.**) (e.g. "The reader is Female, Asian and 28") and (iii) a description of a single characteristic, for which we specifically test **ethnicity** (e.g. "Asian"). Additionally, inspired by recommender system approaches (Shin et al., 2023), we explore (iv) representations of each annotator's label history (**ann. behavior**) as the set of prior WiC instances they labeled as hateful (drawn from the training set). All textual inputs are processed using the same ModernBERT encoder, with the [CLS] token representation used as the embedding. For behavior-based inputs, which consist of a list of texts, the [CLS] representations are averaged to produce a single embedding.

**Feature Weighted Classifier (FWC)** To integrate the auxiliary embeddings with the primary text representation, we adopt a feature-wise learnable weighting scheme. Each auxiliary embedding is assigned a scalar weight (learned during training) that determines its contribution. The weighted auxiliary vectors are then concatenated with the primary text embedding and passed into the classifier. The classifier is a single-layer multilayer perceptron (MLP) comprising a linear transformation, ReLU activation, dropout regularization, and a final linear layer mapping to the output classes.

**Experimental setup** We evaluate ten model configurations: three using only primary inputs (i.e. the WiC and/or its definition), and seven that additionally incorporate auxiliary annotator information. Model predictions are generated for each individual annotation in the HateWiC dataset using a 10-fold cross-validation framework. Each fold follows a fixed 80-10-10 split into training, validation, and testing sets. Further implementation details, including libraries, hyperparameters, and hardware specifics, are provided in Appendix C.

### 5.2 Evaluation

Our goal is to assess how closely computational models capture human variation in annotation for the HateWiC task. In the previous section, we statistically analyzed a range of linguistic and annotator-specific features to understand their influence on human agreement. Here, we evaluate

whether models can replicate these patterns by analyzing their predictions of individual annotations (typically three per sentence), with and without annotator-specific information as auxiliary input. In the latter case, models simulate predictions from annotators by conditioning on annotator identity.

**Prediction agreement** To quantify how closely the model's predictions resemble human annotation variation in terms of inter-annotator agreement measured through Krippendorff's alpha ($\alpha$), we define an **Agreement Alignment** score as:

$$\mathbf{AA} = 1 - |\alpha_{\text{model}} - \alpha_{\text{human}}|$$

Here, $\alpha_{\text{model}}$ is computed over the model's predicted annotations and thus reflects the model's variation across simulated annotators. $\alpha_{\text{human}}$ is the alpha from actual human annotations. The score ranges from 0 to 1, with higher values indicating that the degree of variation in the model's predictions more closely matches the degree of variation observed in human annotations.

**Agreement patterns** To assess whether models go beyond surface-level agreement and replicate deeper variation patterns, we examine whether they reproduce the same effects of linguistic and annotator variables on label variation as observed in human data. Specifically, we conduct the same statistical tests (§4), replacing human annotations with model predictions. Variation is again treated as a binary variable (agree or disagree) based on whether each individual model prediction aligns with the model-level majority vote. This mirrors the human data procedure, where individual annotations were compared to the human majority.

Using the same set of ten linguistic and annotator features listed in Table 2, we examine both main effects of individual features (10 effects) and interactions between feature pairs, i.e. 45 effects from all possible pairwise combinations. To quantify how closely a model replicates variation patterns, we compute the **Relative Pattern Alignment (RPA)** score between human and model effect sizes (measured using Cramér's V) across all $n$ effects, which we define as:

$$\mathbf{RPA} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\text{effect}_{\text{human},i} - \text{effect}_{\text{model},i}}{\text{effect}_{\text{human},i}}\right|$$

We normalize each error by the magnitude of the corresponding human effect size to accommodate the small magnitude of Cramér's V and to prevent larger effects from disproportionately influencing the score. The final metric is inverted so that higher RPA values indicate stronger alignment between the model's and humans' variation patterns.

**Prediction accuracy** Finally, we directly compare the model predictions to individual human annotations, following traditional evaluation practices. For each model, we report **accuracy** across all instances.

## 6 Results & Discussion

We present results for all ten model configurations in Table 3, which vary in terms of their input features: (i) primary input only, (ii) primary input with *id-based* annotator embeddings, and (iii) primary input with *text-based* annotator embeddings.

| FWC config. | model input | AA | RPA | Acc. |
|---|---|---|---|---|
| primary only | WiC | 0.452 | 0.000 | 0.650 |
| | WikDef | 0.452 | 0.000 | 0.671 |
| | WiC + WikDef | 0.452 | 0.000 | 0.700 |
| + aux. (*id-based*) | WiC + ann. id | 0.670 | 0.620 | 0.664 |
| | WikDef + ann id | **0.732** | 0.632 | 0.682 |
| | WiC + WikDef + ann. id | 0.638 | **0.658** | **0.704** |
| + aux. (*text-based*) | WiC + ann. id | 0.516 | 0.567 | 0.656 |
| | WiC + profile descr. | 0.576 | 0.557 | 0.654 |
| | WiC + ethnicity | 0.501 | 0.539 | 0.654 |
| | WiC + ann. behavior | 0.452 | 0.000 | 0.654 |

Table 3: Agreement Alignment score, Relative Pattern Alignment score and accuracy for the different model configurations compared to the human annotation data.

### 6.1 Prediction agreement

Models that process only primary text naturally produce identical predictions across simulated annotators for each instance. This results in perfect inter-annotator agreement ($\alpha_{\text{model}} = 1$). Consequently, they score lowest on Agreement Alignment (AA = 0.452), as they fail to reproduce the human variation in annotations. In contrast, models that incorporate auxiliary annotator information, particularly those with *id-based* embeddings, exhibit lower agreement rates. This indicates that simulated annotators produce diverging predictions on the same primary input, mimicking the variation observed in human annotations.

Text-based auxiliary inputs result only in modest improvements over primary-only baselines and underperform compared to *id-based* embeddings. For instance, using *text-based* annotator ids yields an AA of 0.516, whereas the corresponding *id-based* configuration achieves 0.670. These differences

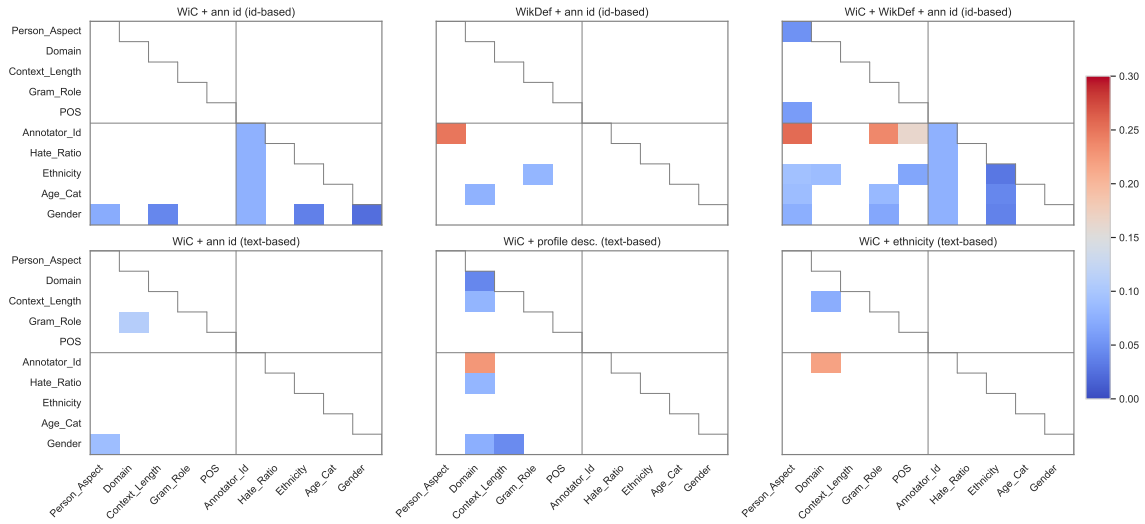


Figure 3: Heatmaps of Cramer's V effect sizes showing both individual (along the diagonal) and pairwise associations of features with model prediction agreement for different FWC model configurations (named after their inputs). The upper triangles (above the diagonal) as well as non-significant ($p > 0.05$) interaction effects are masked.

might originate from the fact that *id-based* embeddings are jointly trained, letting the model distinguish the annotators in a more clear-cut manner, whereas *text-based* inputs rely on static representations from a pre-trained encoder, limiting their influence on the model's decision making. Notably, the WiC + ann. behavior model maintains perfect inter-annotator agreement ($\alpha_{\text{model}} = 1$), suggesting that the behavior representations do not inject any variation into model predictions. A possible explanation is that each annotator's behavior embedding is a fixed average of the hateful sentences they labeled, which may smooth out fine-grained differences and lack strong signals to distinguish annotators.

Overall, these findings suggest that conditioning on annotator identity introduces label variation, but the way this auxiliary input is provided affects the extent of this variation. Yet, in general, models underestimate the magnitude of variation observed in human annotations.

### 6.2 Agreement patterns

While Agreement Alignment quantifies whether models produce human-like variation in an aggregated manner, it does not capture *how* that variation arises. To probe this, we analyze Relative Pattern Alignment (RPA), which measures how well a model replicates the internal structure of human variation. High AA does not always translate to high RPA, indicating that the variation in human data and model predictions might originate from different instances. For example, while the model with WikDef + ann. id has the highest AA (0.732), the configuration with combined inputs (WiC + WikDef + ann. id) achieves the best RPA (0.658). These results reveal that surface-level agreement can be misleading, since it does not guarantee alignment with the internal structure of human variation.

Figure 3 visualizes feature association patterns for each of the six models, restricted to those exhibiting variation in their predictions ($\alpha_{\text{model}} < 1$). It displays for each model a heatmap of Cramer's V effect sizes showing both individual and pairwise associations of features with prediction agreement. The human annotation data showed a diverse range of significant effects (48 out of 55 tested), including interactions between annotator features, linguistic features, and cross-type combinations. Among these, the latter were particularly prominent. Models generally captured far fewer significant effects and vary widely in their replication of human-like effect structures. A key distinction emerges in the types of feature interactions that models are able to replicate. The best model in terms of RPA (WiC + WikDef + ann. id) captures numerous significant effects spanning all three interaction types. In contrast, only two significant effects were identified for the model with WiC + ann id (*text-based*) inputs, none involving annotator × annotator interactions.

Overall, these findings show the importance of not just measuring agreement rates, but also systematically analyzing the patterns of variation, which can offer a more fine-grained view of how model predictions reflect the structure of human annota-

tion behavior.

### 6.3 Prediction performance

Across all configurations, predictive accuracy remains relatively stable (0.65–0.70). The highest accuracies are observed for models using semantic-rich inputs, i.e. including both sentence context (WiC) *and* definitions (WikDef) as inputs. This highlights the importance of linguistic information for predicting individual annotations and aligns with our earlier findings on the role of linguistic features in human annotations. In addition, models that best capture human-like variation do not necessarily predict individual labels more accurately. For instance, although the WikDef + ann. id model exhibits strong AA (0.732) and RPA (0.632), its accuracy (0.682) is only marginally better than primary-only models. These findings suggest that optimizing for predictive accuracy and optimizing for alignment with human variation may constitute distinct modeling objectives that warrant separate consideration in model development.

## 7 Conclusion

In this paper we demonstrated that the variation in interpretation of hateful word meaning is not merely a function of *who* the annotator is or *what* is being annotated, but of the interaction between the two. Through empirical analysis of the HateWiC dataset, we showed that both linguistic properties of the target word in context and annotator characteristics shape interpretive variation. Our evaluation of model alignment with human variation further reveals that although models that incorporate annotator-specific information introduce human-like variation at a surface level, they still underestimate the magnitude of variation observed in human annotations and generally fail to represent the internal structure of variation. In conclusion, our findings show that capturing human interpretive variation requires modeling the interaction between annotators and linguistic content, and that surface-level agreement or predictive accuracy alone does not ensure true alignment with human variation.

## Limitations

Alongside its contributions, this study has several limitations that should be acknowledged:

**Binary operationalization.** Our analysis relies on binary categorizations for hatefulness (hateful vs. not hateful) and annotator agreement (agree vs. disagree with majority). While this simplifies modeling and interpretation, it risks oversimplifying the complexity of human judgments. Future work could explore multi-class or continuous scales to capture finer distinctions in hatefulness and annotation variation.

**Feature selection & categorization.** The linguistic and annotator features included in our study, although carefully chosen to cover key linguistic and annotator dimensions, represent a subset of potentially relevant factors. Additionally, some features were either provided in broad categories or grouped during analysis to facilitate reliable statistical analysis. Other linguistic phenomena, richer annotator identity information and more refined categorizations might further explain variation patterns.

**Label variation as interpretation variation.** We interpret label variation among annotators as indicative of variation in meaning interpretation. While this is a reasonable assumption, other sources of disagreement, such as sloppy annotations or uncertainty, cannot be fully ruled out (e.g. Sandri et al. (2023)). Incorporating complementary data such as annotator confidence ratings or qualitative feedback could strengthen this.

**Automatic parsing.** The Grammatical Role feature was derived using automatic dependency parsing (SpaCy) without additional validation tailored to the specific dataset. While SpaCy generally offers robust performance, parsing errors could introduce noise in the linguistic feature set. Dataset-specific parser evaluation could improve feature reliability in future analyses.

**Data size and imbalance.** Some feature categories have limited observations, restricting the use of complex models like mixed-effects regression with random intercepts for annotators. These models treat each subcategory as a separate binary feature which requires enough data per subcategory to produce reliable estimates of variation and interaction effects. Due to this, we relied on Chi-squared tests and effect size measures better suited to the dataset. Larger, more balanced data would enable exploring richer feature effects.

**Limited modeling diversity.** The modeling component of this study focused on one type of architecture (ModernBERT-based encoder models with auxiliary feature integration). While this design allowed us to systematically evaluate the contribution

of annotator information within a controlled setup, it does not explore the full range of potentially useful architectures. Future work could broaden this scope to assess generalizability across modeling paradigms.

## Ethics Statement

Our work builds upon the HateWiC dataset by enriching it with additional linguistic annotations and computational analyses. Apart from the supplementary linguistic annotations (see also Appendix A), no new human annotations were collected for this research beyond what is already available in HateWiC, and no personally identifying information was processed or used. Where annotator identity is used for modeling purposes, it is limited to anonymous identifiers that cannot be traced to real individuals. We recognize that demographic categories such as ethnicity, gender and age provide only a limited representation of individual identity. These features are used here solely to explore variation in annotator interpretation and not to make generalizations about any group.

Given the sensitive nature of hate-related content, we have taken care to conduct our analyses and reporting in a manner that avoids harm. The focus of our work is on variation in interpretation rather than the endorsement or rejection of any specific viewpoint. Our goal is to improve understanding of the variation inherent to such subjective annotation tasks, in order to support the development of computational methods that better account for subjective variation and promote fairness in NLP applications.

## Acknowledgments

The authors acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of North Rhine-Westphalia (Germany) and by the project "Dealing with Meaning Variation in NLP", which is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant (NGF.1607.22.002).

## References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 187–205, Miami, Florida, USA. Association for Computational Linguistics.

Shayan Alipour, Indira Sen, Mattia Samory, and Tanushree Mitra. 2024. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness.

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752,

Dublin, Ireland. Association for Computational Linguistics.

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

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Aldo Frigerio and Maria Paola Tenchini. 2019. Pejoratives: a classification of the connoted terms. *Rivista Italiana di Filosofia del Linguaggio*, 13(1).

Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2023. Identifying slurs and lexical hate speech via lightweight dimension projection in embedding space. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 278–289, Toronto, Canada. Association for Computational Linguistics.

Sanne Hoeken, Sina Zarrieß, and Özge Alacam. 2024. Hateful word in context classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 172–186, Miami, Florida, USA. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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

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.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12493–12509, Bangkok, Thailand. Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2022. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1):155–188.

Barbara Partee et al. 1984. Compositionality. *Varieties of formal semantics*, 3:281–311.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

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.

Christopher Potts. 2007. The expressive dimension. *Theoretical Linguistics*, 33(2):165–198.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future: 17th International Conference, IConference 2022, Virtual Event, February 28 – March 4, 2022, Proceedings, Part I*, page 425–444, Berlin, Heidelberg. Springer-Verlag.

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.

Asad Sayeed. 2013. An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 691–696, Atlanta, Georgia. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Kyuyong Shin, Hanock Kwak, Wonjae Kim, Jisu Jeong, Seungjae Jung, Kyungmin Kim, Jung-Woo Ha, and Sang-Woo Lee. 2023. Pivotal role of language modeling in recommender systems: Enriching task-specific and task-agnostic representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1146–1161, Toronto, Canada. Association for Computational Linguistics.

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. Publisher Copyright: © 2021 AI Access Foundation. All rights reserved.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France. European Language Resources Association.

# A Data

We retrieved the HateWiC dataset upon request which is available for research purposes, licensed under CC BY-NC 4.0.

## A.1 Sense-level annotation

The annotation task was conducted on Wiktionary definitions from the HateWiC dataset, comprising nearly 1,900 instances. Each instance was annotated with two categorical labels: one for semantic Domain and one for Person Aspect. The Domain label captures the conceptual domain of the term, provided that its part of speech is a noun; otherwise, it is labeled as *NotNoun*. The Person Aspect label identifies what aspect of a person the sense pertains to, and is only applied if the term refers to a person; otherwise, it is labeled as *NotPerson*.

The Domain taxonomy includes the following categories: *Person*, *Animal*, *Artefact*, *Body part*, *Disease*, *Food*, *Plant*, *Supernatural being*, *Ambiguous* and *Other*. The Person Aspect taxonomy includes: *Personality/behavior*, *Ethnicity/nationality*, *Health/disability*, *Intelligence*, *Profession*, *Politics/ideology*, *Appearance*, *Gender/sexuality*, *Kinship/social*, *Social class*, *Age* and *Unspecified*.

Full annotation guidelines, including definitions of each category, are available in our GitHub repository. The annotation was carried out by two annotators with expertise in linguistics: Annotator 1 (author) is a PhD student in Computational Linguistics and Annotator 2 is a student in English and Computational Linguistics. Inter-annotator agreement, measured using Cohen's kappa, was $\kappa = 0.832$ for the Domain annotations and $\kappa = 0.764$ for the Person Aspect annotations. Annotator 2's annotations served as validation, with Annotator 1 providing the authoritative judgment when consensus was not reached.

### A.2 Grammatical Role extraction

We implemented a custom pipeline using the `spacy` nlp library with the `en_core_web_sm` model. To locate predefined (multiword) terms within sentences, we used spacy's `PhraseMatcher`, configured to match on the lemmatized form of the target terms (using spacy's built-in lemmatizer). If no exact lemmatized match was found, approximate string matching was performed using the `rapidfuzz` library, leveraging the Levenshtein similarity ratio. Candidate noun chunks in each sentence were compared to the expected lemmatized term, and the highest-scoring match above a fuzzy similarity threshold of 85 was selected. For both exact and approximate matches, the syntactic role of the term was determined by extracting the dependency label (`dep_`) of the syntactic head of the matched span. Processing was parallelized using spacy's `nlp.pipe` API with a batch size of 50.

After extracting the dependency parsing tags using SpaCy for the target terms in the texts, we mapped them to a coarser categorization based on guidelines provided in https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md. The coarse-grained categories of Grammatical Roles are: *subject*, *object*, *nominal*, *adverbial*, *preposition*, *coordination*, *root*, *compoundword*, *complement* and *miscellaneous*.

### A.3 Annotator Hatefulness Ratio

We computed each annotator's Hatefulness Ratio, defined as the proportion of instances they labeled as hateful across the dataset, i.e.:

$$H_a = \frac{N_a^{(h)}}{N_a}$$

where $H_a$ denotes the Hatefulness Ratio of annotator $a$, $N_a^{(h)}$ is the number of instances annotator $a$ labeled as hateful, and $N_a$ is the total number of instances annotated by $a$.

## B Empirical Analysis

Inter-annotator agreement was computed using the `krippendorff` package. Furthermore, we conducted two types of statistical analyses. Feature association testing was carried out using chi-squared tests of independence via the `scipy.stats` package. For handling numerical variables, we applied quantile-based binning to create discrete categories. This was achieved using the `qcut` function from the `pandas` library.

For the analysis visualized in Figure 4, ordinary least squares (OLS) regression was applied using the `OLS` method from the `statsmodels.api` module. We included interaction and polynomial terms using `PolynomialFeatures` from `sklearn.preprocessing` and computed the coefficient of determination ($R^2$) with `sklearn.metrics.r2_score`.

All data visualizations were produced with `matplotlib.pyplot` and `seaborn`.

## C Computational Modeling

All modeling experiments were implemented using the PyTorch framework. Text representations were obtained using a pre-trained transformer model. More specifically, initialized with the `'answerdotai/ModernBERT-base'` checkpoint via the `transformers` library. Model training was performed with the Adam optimizer using a learning rate of $2 \times 10^{-5}$ and a batch size of 32 for both training and evaluation. The training process was conducted over 3 epochs using a fixed random seed of 56 to ensure reproducibility. Classification performance was evaluated using cross-entropy loss and accuracy computed with `sklearn.metrics`. All experiments were executed on a single NVIDIA RTX A6000 GPU using CUDA acceleration.

## D Additional Results

### D.1 Main effect of Hatefulness Ratio

An additional illustration of the directions of feature effects is provided in Figure 4. The figure plots individual annotators' hateful labeling tendency (Hatefulness Ratio) against their annotation

agreement ratio with the majority, which allows for more concrete interpretation of this measured main effect as presented in Table 2. Unlike earlier analyses, which relied on binned categories, this figure presents the continuous relationship between these variables. The relationship appears weakly quadratic, with lower agreement visible at both extremes of Hatefulness Ratio. As expected, annotators who rarely or frequently label instances as hateful tend to deviate more often from the majority decision, while those with moderate Hatefulness Ratios agree more frequently. Especially for these annotators, incorporating individual labeling behavior may improve models of annotation variation.

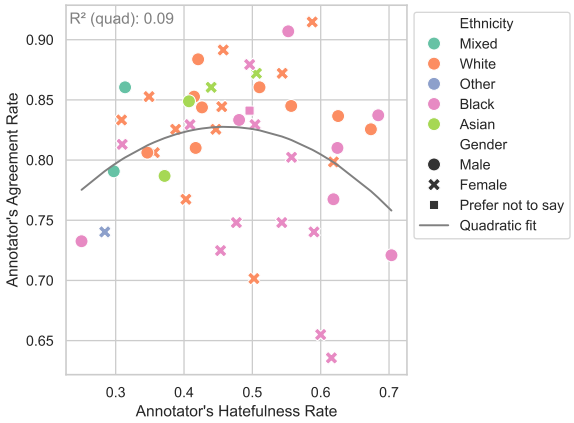


Figure 4: Annotator's hatefulness proportion (i.e. how much of their annotations is hateful) against agreement ratio (i.e. how much of their annotations is the majority vote). Each datapoint represents one annotator.

## D.2 Contingency tables

For each feature, the frequency counts that underlie the statistical analyses in Table 2 are reported in Tables 4 until 13.

| Person Aspect | agree | disagree |
|---|---|---|
| **Age** | 77 | 16 |
| **Appearance** | 307 | 41 |
| **Ethnicity/nationality** | 480 | 80 |
| **Gender/sexuality** | 510 | 74 |
| **Health/disability** | 171 | 17 |
| **Intelligence** | 380 | 52 |
| **Kinship/social** | 164 | 14 |
| **NotPerson** | 4271 | 688 |
| **Personality/behavior** | 2183 | 473 |
| **Politics/ideology** | 772 | 185 |
| **Profession** | 228 | 43 |
| **Social class** | 49 | 10 |
| **Undecided** | 47 | 10 |
| **Unspecified** | 177 | 18 |

Table 4: Frequencies for individual annotations by Agreement with the majority vote and Person Aspect

| Domain | agree | disagree |
|---|---|---|
| **Ambiguous** | 191 | 28 |
| **Animal** | 248 | 40 |
| **Artefact** | 579 | 50 |
| **Body part** | 212 | 46 |
| **Disease** | 213 | 43 |
| **Food** | 110 | 15 |
| **NotNoun** | 2111 | 377 |
| **Other** | 802 | 120 |
| **Person** | 6786 | 1107 |
| **Plant** | 82 | 12 |
| **Super natural being** | 33 | 7 |

Table 5: Frequencies for individual annotations by Agreement with the majority vote and Domain

| Context Length | agree | disagree |
|---|---|---|
| **3-14** | 2667 | 341 |
| **14-23** | 2494 | 374 |
| **23-35** | 2478 | 512 |
| **35-176** | 2483 | 487 |

Table 6: Frequencies for individual annotations by Agreement with the majority vote and Context Length

| Grammatical Role | agree | disagree |
|---|---|---|
| **adverbial** | 457 | 55 |
| **complement** | 415 | 73 |
| **compoundword** | 704 | 125 |
| **coordination** | 687 | 125 |
| **miscellaneous** | 101 | 18 |
| **nominal** | 1078 | 188 |
| **not_found** | 196 | 48 |
| **object** | 2897 | 486 |
| **preposition** | 1073 | 373 |
| **root** | 600 | 91 |
| **subject** | 1246 | 297 |

Table 7: Frequencies for individual annotations by Agreement with the majority vote and Grammatical Role

| POS | agree | disagree |
|---|---|---|
| **adjective** | 848 | 159 |
| **adverb** | 237 | 30 |
| **interjection** | 47 | 5 |
| **noun** | 7688 | 1421 |
| **phrase** | 3 | 0 |
| **proper noun** | 94 | 11 |
| **verb** | 1193 | 210 |

Table 8: Frequencies for individual annotations by Agreement with the majority vote and POS

| Hate Ratio | agree | disagree |
|---|---|---|
| **0.25-0.4** | 2632 | 480 |
| **0.4-0.48** | 2668 | 482 |
| **0.48-0.56** | 2577 | 382 |
| **0.56-0.7** | 2418 | 535 |

Table 9: Frequencies for individual annotations by Agreement with the majority vote and Hate Ratio

| Ethnicity | agree | disagree |
|---|---|---|
| **Asian** | 889 | 142 |
| **Black** | 3883 | 810 |
| **Mixed** | 639 | 110 |
| **Other** | 191 | 66 |
| **White** | 4511 | 675 |

Table 10: Frequencies for individual annotations by Agreement with the majority vote and Ethnicity

| Age Category | agree | disagree |
|---|---|---|
| **20-29** | 3657 | 1255 |
| **30-39** | 2790 | 960 |
| **40-49** | 210 | 20 |
| **50-59** | 208 | 40 |
| **60+** | 213 | 32 |

Table 11: Frequencies for individual annotations by Agreement with the majority vote and Age Category

| Gender | agree | disagree |
|---|---|---|
| **Female** | 5410 | 1086 |
| **Male** | 4467 | 777 |
| **Prefer** | 217 | 27 |

Table 12: Frequencies for individual annotations by Agreement with the majority vote and Gender

| Annotator Id | agree | disagree |
|---|---|---|
| **annotator_1** | 261 | 33 |
| **annotator_10** | 213 | 32 |
| **annotator_13** | 181 | 32 |
| **annotator_14** | 213 | 32 |
| **annotator_16** | 198 | 59 |
| **annotator_18** | 193 | 39 |
| **annotator_19** | 187 | 60 |
| **annotator_2** | 228 | 24 |
| **annotator_22** | 215 | 43 |
| **annotator_23** | 217 | 27 |
| **annotator_24** | 220 | 26 |
| **annotator_25** | 216 | 31 |
| **annotator_26** | 169 | 31 |
| **annotator_28** | 219 | 29 |
| **annotator_30** | 217 | 29 |
| **annotator_31** | 222 | 17 |
| **annotator_34** | 215 | 41 |
| **annotator_35** | 207 | 37 |
| **annotator_36** | 191 | 53 |
| **annotator_37** | 226 | 28 |
| **annotator_39** | 193 | 42 |
| **annotator_4** | 206 | 51 |
| **annotator_42** | 214 | 38 |
| **annotator_44** | 222 | 35 |
| **annotator_47** | 191 | 66 |
| **annotator_5** | 225 | 27 |
| **annotator_53** | 208 | 36 |
| **annotator_55** | 164 | 78 |
| **annotator_56** | 198 | 18 |
| **annotator_58** | 225 | 28 |
| **annotator_59** | 220 | 38 |
| **annotator_6** | 189 | 39 |
| **annotator_60** | 222 | 33 |
| **annotator_62** | 234 | 23 |
| **annotator_63** | 186 | 67 |
| **annotator_64** | 209 | 44 |
| **annotator_65** | 209 | 26 |
| **annotator_66** | 230 | 28 |
| **annotator_69** | 213 | 40 |
| **annotator_71** | 213 | 32 |
| **annotator_74** | 214 | 44 |
| **annotator_75** | 203 | 50 |
| **annotator_77** | 215 | 36 |
| **annotator_78** | 208 | 49 |
| **annotator_79** | 218 | 37 |
| **annotator_8** | 216 | 35 |
| **annotator_83** | 204 | 45 |
| **annotator_85** | 236 | 21 |

Table 13: Frequencies for individual annotations by Agreement with the majority vote and Annotator Id