Eyes Don't Lie: Subjective Hate Annotation and Detection with Gaze

Özge Alaçam[1,2], **Sanne Hoeken**[1], and **Sina Zarrieß**[1]
[1]Computational Linguistics, Department of Linguistics, Bielefeld University, Germany
[2]Center for Information and Language Processing, LMU Munich, Germany
{oezge.alacam,sanne.hoeken, sina.zarriess}@uni-bielefeld.de

Abstract

Hate speech is a complex and subjective phenomenon. In this paper, we present a dataset (GAZE4HATE) that provides gaze data collected in a hate speech annotation experiment. We study whether the gaze of an annotator provides predictors of their subjective hatefulness rating, and how gaze features can improve Hate Speech Detection (HSD). We conduct experiments on statistical modeling of subjective hate ratings and gaze and analyze to what extent rationales derived from hate speech models correspond to human gaze and explanations in our data. Finally, we introduce MEANION, a first gaze-integrated HSD model. Our experiments show that particular gaze features like dwell time or fixation counts systematically correlate with annotators' subjective hate rating, and improve predictions of text-only hate speech models.

1 Introduction

Hate speech is a real threat that harms individuals, groups, and societies in a profound way. Even though research in NLP has developed many different datasets and models for HSD (Poletto et al., 2021), the accurate modeling of hate speech is far from being solved (Ocampo et al., 2023; Röttger et al., 2021). One of the key challenges in this area is that the definition and annotation of hate speech are highly complex and subjective, depending on the topic and domain of hate as well as on the individual annotators' backgrounds and biases (Waseem and Hovy, 2016; Abercrombie et al., 2023; ElSherief et al., 2018; Kovács et al., 2021). This combines with the fact that state-of-the-art HSD models are typically designed as black-box neural models that are well-known to pick up superficial, dataset-dependent patterns rather than learning a generalizable model of the underlying task. Therefore, it is still an open question of how to handle subjective variation in human annotations and detection of hate speech.

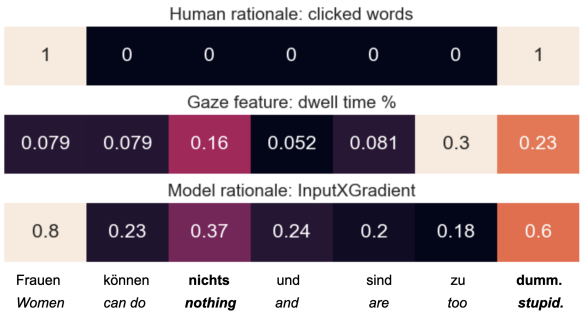


Figure 1: Heatmaps for a human rationale, gaze feature and model rationale for a hateful sentence from GAZE4HATE

This paper contributes a new dataset (GAZE4HATE) that provides gaze and annotations from hate speech annotators, illustrated in Figure 1. We recorded the eye movements of annotators while they read statements, which were carefully controlled and constructed. This was followed by the annotation of hatefulness. Annotators' gaze provides us with an extremely rich signal of the subjective cognitive processes involved in human hate speech evaluation while reading. In this paper, we explore whether subjective hatefulness rating can be predicted by the gaze of an annotator, and whether gaze features can be used to evaluate and improve HSD models.

Generally, the NLP community has recently started to leverage eye-tracking data as a means of analyzing the internal mechanisms in transformer language models as elaborated on in Section 2.1. To the best of our knowledge, however, there is no available dataset of human reading of hate speech. Other work along these lines has adopted so-called rationale annotations, where annotators mark text spans that they consider indicative of their labeling decisions (e.g. DeYoung et al. (2020); Mathew et al. (2021)). These rationales can be used to measure the plausibility and explainability of model decisions, by testing whether model-internal weights

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 187–205
November 12-16, 2024

and gradients correlate with or even predict these human rationales (Atanasova et al., 2020). Yet, to date, it is unclear how rational annotations compare to gaze signals recorded during plain reading for the task of hate speech classification. Our GAZE4HATE data closes this gap, as our annotators did not only rate texts for hatefulness but also annotated token-level rationales for their ratings. Figure 1 shows an example that illustrates human gaze and rationales aligned with a model's rationale.

Our analyses and experiments center around the following research questions:

- RQ1 Do gaze features provide robust predictors for subjective hate speech annotations?
- RQ2 How do gaze features correlate with human and model rationales?
- RQ3 Are gaze features useful for enriching LMs for HSD?

We address the first question by conducting statistical modeling on our collected eye-tracking and annotation data (Section 4). To answer the second question, we evaluate a range of existing HSD models on our data, comparing models' and humans' rationales to human gaze (Section 5). Section 6 presents the MEANION model, which integrates text-based HSD with gaze features. In sum, our experiments show that particular gaze features like dwell time or fixation counts systematically differ with respect to annotators' subjective hate ratings. Models' rationales, however, correlate more with explicit, annotated rationales than with annotator gaze. Finally, in some settings, adding gaze features improves predictions of text-only hate speech models more than human rationales do.

2 Related Work

2.1 Eyetracking Data in NLP

In work on testing the cognitive plausibility of attention-based transformer language models, human gaze is a very relevant indicator of readers' cognitive processes and a valuable source of evaluation data (Das et al., 2016; Malmaud et al., 2020; Sood et al., 2020; Hollenstein and Beinborn, 2021; Eberle et al., 2022; de Langis and Kang, 2023). Unfortunately, the collection of eyetracking data is costly and existing task-specific datasets are small and scarce (de Langis and Kang, 2023). Our work contributes to enriching the landscape of available NLP-tailored eyetracking datasets.

Previous studies on using gaze to extend NLP models usually focus on a few high-level gaze features (Barrett et al., 2016; Long et al., 2019; Eberle et al., 2022), with some exceptions (Mishra et al., 2017; Hollenstein et al., 2019; Alacam et al., 2022). As one of the most commonly used group of gaze features in NLP, fixations measure the pause of the eye movement on an area of the visual field, and are strongly associated with visual intake (Rayner, 1998; Kowler, 2011; Skaramagkas et al., 2021). However, reading hateful text also involves intense emotions (e.g. feeling empathy, being the target of the hate speech). Little NLP work has been done on emotion-related eye movements such as *pupil dilation*, which is associated with emotional and cognitive arousal (Bradley et al., 2008). Our work considers a range of gaze features and compares their predictive power for subjective hate ratings. Furthermore, gaze features are commonly preprocessed in non-trivial ways, e.g. by aggregating all token-level features or arranging them in a token-based discretized sequence as in the above-mentioned studies. We adopt such a simple token-based preprocessing for our MEANION model, and leave exploration of more advanced architectures such as time series-based gaze transformers (Alacam et al., 2022) for future work.

2.2 Explainability

To assess whether models attend to relevant parts of an input, various explanation and rationale extraction methods have been developed, e.g., model simplification methods (Ribeiro et al., 2016), gradient-based techniques (Simonyan et al., 2014; Sundararajan et al., 2017), perturbation-based methods (Zeiler and Fergus, 2013) and Shapley-based methods (Shapley, 1953). The work of Atanasova et al. (2020) evaluates different methods for text classification models, concluding that "the gradient-based explanations perform best across tasks and model architectures". Yet, the 'best' method *highly* depends on the dataset/task, model, and diagnostic property used for evaluation. In this study, we evaluate a selection of explanation methods for hate speech classification, which has not been attempted before. We do so not only on human annotations of salient tokens (as e.g. Atanasova et al. (2020) did) but also on human gaze measurements.

2.3 Hate Speech and Subjectivity

Since the advent of research on hate speech detection (HSD), the reliable annotation of hate in texts

has been recognized as a notorious issue (Waseem, 2016; Schmidt and Wiegand, 2017). Still, HSD is often modeled with text classifiers, trained and fine-tuned on ground-truth annotations and benchmarks (Davidson et al., 2017; Basile et al., 2019; Zampieri et al., 2019). Recent approaches and shared tasks, though, shifted the focus to specific domains of hate such as sexism (Kirk et al., 2023) as well as explainable HSD (Mathew et al., 2021; Pavlopoulos et al., 2022; ElSherief et al., 2021). Röttger et al. (2021) present the HateCheck benchmark, which is composed of linguistically controlled functional tests designed to systematically assess language understanding in hate speech models. Davani et al. (2022) take some first steps in dealing with disagreements between annotators in HSD and compare the prediction of majority vote vs. individual labels. Similarly, Wojatzki et al. (2018) compare hate speech annotations of female and male annotators on hateful statements about women.

Furthermore, there is an emerging research that explores the contribution of injecting annotators' demographics and preferences along with the annotated text (Kanclerz et al., 2022; Fleisig et al., 2023). The results of these studies indicate that demographic information is a successful predictor for annotators' ratings on the sentence-level hate speech. Furthermore, Hoeken et al. (2024) shows that annotator's demographics are also useful for predicting subjective annotations at the lexical level i.e. predicting hateful words in context.

Our collection of annotator gaze provides a new direction for tackling the issues of explainability and subjectivity in an integrated fashion.

3 GAZE4HATE Dataset

We collected a hate speech annotated dataset that provides information from three different sources: hatefulness ratings of text w.r.t. gender, eye movements during plain readings of the statements, and explicit rationales marked by annotators. In this section, we explain the design of the dataset.

3.1 Data and Sentence Selection

To obtain a dataset for systematic analysis of hate speech understanding in models, and of subjective differences between annotators and their gaze, we opted for a carefully controlled set of constructed items, similar to Röttger et al. (2021). As is common in eyetracking studies in linguistics, we design our items as minimal pairs: we first collect a set of "seed" hateful statements. Within these statements, we manipulate specific tokens that change the hatefulness of the statement and turn it into a neutral or even positive statement. Furthermore, we consider (i) items that express hate explicitly, through direct lexical cues, and (ii) items where the expression of hate is implicit and results from the social meaning of the sentence as a whole. These conditions roughly correspond to the explicit vs. implicit derogation category in Röttger et al. (2021)'s HateCheck taxonomy.

As an example, consider the hateful statement *Women can do nothing and are too stupid* in Table 1. When *women* is replaced with *minions*, the statement is neutral towards women. When changing *nothing* and *stupid* the meaning of the statement even turns positive. This example belongs to the "explicit" condition in our design as it contains hateful lexical cues (e.g. *stupid*). The statement *Women belong in the kitchen* illustrates the "implicit" condition, as none of its words is hateful on its own. Analog to the "explicit" condition, minimal pairs can be constructed, i.e., when changing *women* to *pots*, the statement is not hateful anymore.

We conducted our in-lab experiments in Germany and based the construction of our annotation targets on the German FEMHATE (Wojatzki et al., 2018) dataset that contains 400 statements about women. We have selected 20 sentences from FEMHATE with respect to two criteria; (i) being rated as hateful and (ii) allowing for minimal pair manipulation (for comparable hateful-neutral-positive triples). With minimal pair manipulations, we have created 20 positive statements and 20 neutral (w.r.t gender context) sentences. This subset of 60 sentences constitutes our main manipulation instances. To avoid participants becoming aware of these manipulations, we have introduced control sentences (10 mean and 10 no-hate sentences from the FEMHATE dataset, and 10 new sentences that express hate against men, see Table 1). In total, we have collected eye movements on 90 items, divided into 40 a priori hateful and 50 a priori non-hateful sentences. The selection from the FEMHATE dataset and controlled sentence creations were conducted by one of the authors and the stimuli set was finalized after the validation by two other NLP researchers. For the validation of the a priori-level annotations, we used the definition inspired by Davidson et al. (2017) and the EU

Table 1: GAZE4HATE provides annotated statements about women, divided into 3 conditions (explicit, implicit, control). Explicit and implicit examples figure in minimal pairs where words (in pink and lime) in (a priori) hateful sentences are manipulated to obtain neutral and positive examples. Control examples are not manipulated. **n** = number of items in this category, **P** = number of participants recorded.

Cond.	Example	Translation	A priori label	n	P
Explicit	Frauen können nichts und sind zu dumm.	Women can do nothing and are too stupid.	hate	10	40
	Minions können nichts und sind zu dumm.	Minions can do nothing and are too stupid.	neutral	10	40
	Frauen können alles und sind sehr kreativ.	Women can do everything and are very creative.	positive	10	40
Implicit	Frauen gehören in die Küche.	Women belong in the kitchen.	hate	10	40
	Töpfe gehören in die Küche.	Pots belong in the kitchen.	neutral	10	40
	Frauen gehören in die Regierung.	Women belong in the government.	positive	10	40
Control	Frauen können ohne Kinder glücklich sein.	Women can be happy without children.	neutral (no hate)	10	40
	Spezielle Frauenquoten sind unnötig.	Special quotas for women are unncessary.	hate(unclear/mean)	10	40
	Alle Männer sind Machos.	All men are machos.	hate (men)	10	40

Code of Conduct[1] on hate speech formulated as "*any rude, hurtful, derogatory language that upsets or embarrasses people or groups of people and the extreme form of hate speech incites violence and hatred*".

3.2 Experimental Procedure for Subjective Hate Speech Annotation

Our study follows a within-subject design, i.e. all subjects read and rate all items. Each trial consists of two phases. In the first phase, we record annotator's eye movements while they read the statements. In the second phase, we collect their explicit annotations. We ask participants to rate the statement's hatefulness, to rate their confidence and to mark the words in the statement that contribute to their rating decision. The order of sentences was randomized for each participant.

Participants. 43 university students (native speakers of German) participated in the experiment (32 female, 10 male, 1 non-binary, Mean age = 23.5, SD = 5.3). They were paid or given a course credit to participate. The experiment took approximately 40 minutes for each participant.

Eyetracking Procedure. The stimuli were displayed on an SR Eyelink 1000 Plus eye tracker integrated into a 27" monitor with a resolution of 2560 × 1440. We utilized a total of 94 sentences (including 4 familiarization trials). Each trial began with a drift correction located to the left of the sentence onset location. Then followed the reading phase, in which the participants read the sentence at their own pace. We set a time limit of 20 seconds for the reading task, but the participants were instructed to read as quickly as possible.

Annotation Procedure. The instruction given to the participants is detailed in Appendix A.1. For collecting subjective annotation, we intentionally did not provide a strict hate speech definition to be able to get annotators' interpretation of the statements closest to their personal stance.

First, participants rated the hatefulness of the statement in 1-to-7 Likert Scale (1:very positive, 2:positive, 3:somehow positive, 4:neutral, 5:mean, 6:hateful, 7:extremely hateful). Next, they rated their confidence regarding their rating on a 5-Likert scale (1:not certain, 2:somewhat certain, 3:moderate, 4:certain, 5:very certain). Finally, they annotated the rationale for the decision, by clicking words in the statements that contributed most to their rating. Figure 1 (top) illustrates the rationale annotation.

3.3 Overview

GAZE4HATE provides gaze, hatefulness ratings and rationales for 90 items and 43 participants each summing up to 3870 unique instances of subjective hate ratings[2]. Our dataset is comparable in size to existing eye-tracking datasets like, e.g. (de Langis and Kang, 2023). Figure 2 shows the average subjective hate ratings given by participants for a priori categories. Some sentences were rated differently than their a priori labels (especially a priori positive ones as neutral). The subjective ratings for sentences in other a priori categories also exhibit variations except for the very hateful statements

[1] https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[2] The data and code are publicly available to the research community under a CC-BY-NC 4.0 license at https://gitlab.ub.uni-bielefeld.de/clause/gaze4hate

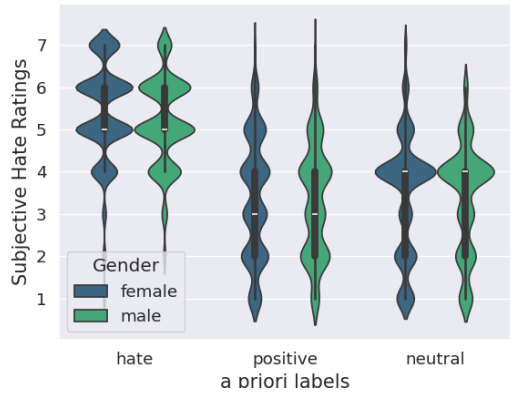


Figure 2: Subjective hate ratings in GAZE4HATE w.r.t. annotators' gender for the a priori labels

(Appendix B.3). These mismatches between the a priori labels and our human ratings once again underline the fact that subjectivity is one of the major challenges in hate speech annotation. Yet, for this study, variation in the annotator's ratings is a feature rather than a bug as it allows us to study subjective hate speech annotations with the help of gaze features, which are highly participant-specific. For the following analysis, we group sentence-based subjective hate ratings provided by users into their hate speech labels (<=3:positive, 4:neutral, >=5:hate).

Train-Test Splits. Sentences from each a priori category were split into three groups (train, validation and test) with a 70:10:20 ratio using 5-fold cross-validation. Each split has instances from each participant, but not from the same sentence.

Preprocessing Gaze Features. Eye movements often show participant-specific patterns and comparing raw gaze features can be misleading. We normalized gaze features with min/max scaling for each participant separately. The description of each feature and pre-processing steps are given in the Appendix A.3.

4 Analysis of Annotators' Gaze

We start with testing whether the gaze parameters show significant differences among the subjective hate categories. We use Anova tests using the OLS library in R on the continuous gaze features. On the categorical gaze features, we utilized Chi-square tests. Multiclass comparison is conducted among hate, neutral and positively rated statements. The binary classification (similar to many existing hate speech classifiers) involves hate and non-hate categories. The non-hate category consists of both neutral and positive statements. For each gaze feature, we checked whether there is a significant main effect of subjective hate categories on the gaze features. Table 2 presents F-scores and significance levels of the above-mentioned statistical tests. The first two columns in the table correspond to measurements on all tokens in the dataset, the last two columns on the right present the results conducted only on the words selected as rationales.

Six out of 13 features consistently show significant differences with high F-score values between the subjective hate ratings for multiclass (hate, neutral, and positive) and for binary comparisons (hate and no hate): FIXATION-COUNT, DWELL-TIME, MAX-FIX-PUPIL-SIZE, MIN-FIX-PUPIL-SIZE, AVERAGE-FIX-PUPIL-SIZE and FIRST-RUN-FIXATION-COUNT. Some features result in low F-score values despite showing significant differences in terms of subjective hate rating. In the following, we remove features that yield low F-scores or non-significant results.

All features that are significant in the multiclass condition are also significant in the binary one, but not the other way around. This indicates that merging neutral and positive categories has a negative impact on the statistical difference. FIXATION-COUNT, DWELL-TIME and FIRST-RUN-FIXATION-COUNT are showing higher F-scores in the binary comparison. Tukey's tests for pairwise comparisons indicate that the differences in the fixation and dwell time originate from the difference between the hate vs. neutral and hate vs. positive conditions, while there is no difference between neutral and positive conditions. On the other hand, differences in the pupil size related parameters originate from difference in neutral conditions to hate and positive conditions without showing a significant difference between the latter two. This also confirms the theory of pupil size being more sensitive to the magnitude of the emotion rather than its polarity (Bradley et al., 2008).

5 HSD Models and rationales

In this Section, we evaluate several hate speech detection (HSD) models on our GAZE4HATE dataset to answer RQ2, which are described in Section 5.1. We not only evaluate classification performance (Section 5.2), but also measure the plausibility and explainability of model decisions by looking into

Table 2: F and Chi-square scores (for continuous and categorical features respectively) of multiclass and binary comparison of subjective hate ratings on (i) all tokens and (ii) rationale tokens

Gaze features (on area-of-interests)	Multiclass all tokens	Binary all tokens	Multiclass rationale tokens	Binary rationale tokens
FIXATION-COUNT	28.01**	49.98**	14.86**	28.51**
DWELL-TIME	25.20**	44.25**	13.38**	24.48**
MAX-FIX-PUPIL-SIZE	31.39**	29.38**	14.11**	16.30**
MIN-FIX-PUPIL-SIZE	42.32**	34.82**	23.80**	20.82**
AVERAGE-FIX-PUPIL-SIZE	37.85**	32.84**	19.05**	19.13**
RUN-COUNT	0.61ns.	0.08ns.	6.30**	6.87*
REG.-IN-COUNT	1.04ns.	2.07ns.	1.57ns.	0.03ns.
REG.-OUT-COUNT	0.32ns.	0.56ns.	0.33ns.	0.63ns.
FIRST-FIX.-DURATION	3.28*	0.19ns.	1.59ns.	0.27ns.
FIRST-RUN-FIXATION	41.49**	54.19**	13.00**	11.47**
REG.-OUT	1.04ns.	2.07ns.	1.57ns.	0.03ns.
REG.-IN	1.61ns.	2.37ns.	3.48ns.	0.13ns.
SKIP	0.32**	0.56ns.	0.33ns.	0.63ns.

Table 3: Overview of the off-the-shelf models for HSD in German tested in this study.

	pretrained model	fine-tuning dataset(s)
deepset	G-BERT (Chan et al., 2020)	GermEval 2018 (Wiegand et al., 2019)
ortiz	G-BERT	HASOC 2019 (Mandl et al., 2019)
aluru	M-BERT	Aluru et al. (2020)
rott	G-BERT	Assenmacher et al. (2021), Demus et al. (2022), Glasenbach (2022)
ml6	G-DistilBERT	GermEval 2018 , GermEval 2021 (Risch et al., 2021), Ross et al. (2017), Bretschneider and Peters (2017), HASOC 2019

the model rationales and compare them with the human rationales and gaze features (Section 5.3).

5.1 Models

We tested five off-the-shelf models from HuggingFace, which we named for reference in the remainder of this paper **deepset**[3], **ortiz**[4], **aluru**[5], **rott**[6] and **ml6**[7], respectively. These models are either German (G) or multilingual (M) BERT-based models finetuned on one or more HSD datasets. Rather than aiming to outperform these models on general-purpose hate speech classification, we selected them as candidates to build upon our multimodal models. A more detailed overview of the models is given in Table 3 and in Appendix C.1.

Based on the performance results of the off-the-shelf models on our dataset (Section 5.2), we took the best-performing model for further finetuning.

rott-hc We finetuned the **rott** model (see Table 3) on the German HateCheck corpus[8] (Röttger et al., 2021), which comprises 3645 crafted sentences, of which 2550 hateful and 509 sentences (hateful and non-hateful) are targeting women. Finetuning details can be found in Appendix C.2.

[3] https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

[4] https://huggingface.co/jorgeortizv/BERT-hateSpeechRecognition-German

[5] https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-german

[6] https://huggingface.co/chrisrtt/gbert-multiclass-german-hate

[7] https://huggingface.co/ml6team/distilbert-base-german-cased-toxic-comments

[8] https://huggingface.co/datasets/Paul/hatecheck-german

Table 4: Classification performance (F1-scores) of the different models on the subjective hate ratings.

	n	deepset	ortiz	aluru	rott	ml6	rott-hc
HATE	*1707*	0.51	0.04	0.00	0.59	0.16	**0.66**
NO HATE	*1909*	0.70	0.70	0.69	0.62	**0.71**	0.70
macro avg	*3616*	0.60	0.35	0.35	0.60	0.44	**0.68**
weighted avg	*3616*	0.61	0.39	0.36	0.60	0.45	**0.68**

5.2 Classification results

We evaluate all models regarding the subjective hate ratings of all individual participants. Both human and model output labels are converted to a binary classification scheme (details in Table 8 in Appendix C.3). It must be emphasized that our task is not to detect a majority-class annotation label. Instead, we aim to detect whether a sentence is perceived as hate by an individual.

The F1-scores results are presented in Table 4. **rott** shows the best performance on detecting HATE sentences (F1 on HATE of 0.59), probably due to the fact that this model is the only one that has deliberately been trained to detect sexist hate speech. Fine-tuning this model further on the HateCheck dataset, resulted in a significant performance increase (the **rott-hc** model shows a macro avg. F1 of 0.68).

5.3 Model rationales

Model rationales for the best performing model (i.e. **rott-hc**) were generated using *Captum* (Kokhlikyan et al., 2020), an open source library built on *PyTorch*. Based on Atanasova et al. (2020), we selected three methods that showed the best results for Transformer-based models on a sentiment classification task: (1) InputXGradient (ℓ2 aggregated), (2) Saliency (ℓ2 aggregated) and (3) Shapley value (sampling) [9].

For each sentence, we extract model rationales for both classes, i.e. a rationale for classifying a sentence as HATE and a rationale for classifying that same sentence as NO HATE. The extracted rationales are then converted from sub-word level (the output level that is inherent to BERT-based models) to word level (aligning with the human rationales), by averaging over multiple sub-word values that constitute a single word.

[9] For the details of the algorithms, please visit *Captum* library: https://captum.ai/docs/algorithms

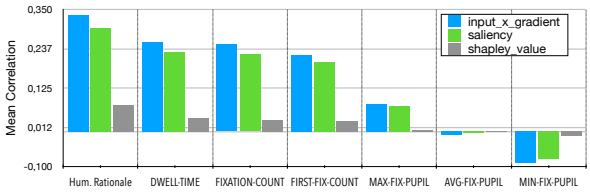


Figure 3: Mean correlation (Pearson's r) between model rationales, human rationales and gaze features.

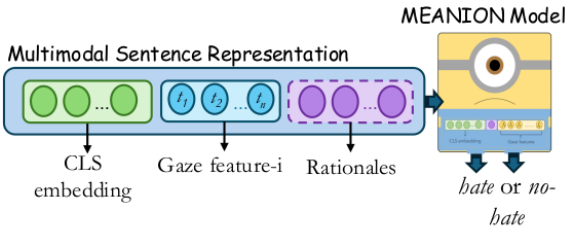


Figure 4: Multimodal sentence representation as input to the MEANION model

For each sentence and annotator, we compare the subjective hate rating (h), human rationale or a gaze feature (f) with a model rationale (r) with respect to class c, where $c = r$. We aggregate correlation values, each calculated as Pearson's r correlation metric between f and r, over all sentences and annotators by taking the mean.

Figure 3 reports mean correlation values of the human rationales and gaze features with the model rationales extracted with different methods (details in Table 9 in Appendix C.4). The six gaze features that showed a significant effect on subjective hate ratings (Table 2) are selected for this analysis. For all human rationales and gaze features, InputXGradient and Saliency rationales show substantially higher correlation than Shapley Value rationales. Additionally, InputXGradient rationales, although less substantial, consistently show higher agreement than Saliency rationales. The variation in agreement among the different gaze features and human rationale show the same pattern for all three rationale methods. Human rationales correlate the highest with model rationales. Among the gaze features, three features, i.e. DWELL-TIME FIXATION-COUNT and FIRST-RUN-FIXATION-COUNT, show a higher correlation (> 0.2) with InputXGradient rationales, while the other three features AVERAGE-FIX-PUPIL-SIZE, MAX-FIX-PUPIL-SIZE and MIN-FIX-PUPIL-SIZE show small to no correlation (between -0.1 and 0.1).

6 MEANION – A Gaze-integrated Baseline Model

In this section, we explore whether gaze features improve pretrained and finetuned models on classifying hate speech (RQ 3). We introduce the first member of our new family of gaze-integrated HSD models (MEANIONS).

6.1 Multimodal Representation

Our MEANION model uses multimodal embeddings that combine three types of embeddings: CLS-token from (L)LMs, token-level gaze features, and rationales as bag-of-words (bow) vector (Figure 4). We trained MLP classifiers using the scikit-learn library[10] on multimodal sentence representations (see Appendix D.3 for the training details).

As changes in eye movement patterns are rather local (e.g. fixation duration increases if the token is unexpected), gaze features for some tokens might be more informative than others for the classification, and averaging over tokens might lose a significant amount of signal. Therefore, we kept the values of each feature for each token in the representation. We first add *text features*. We use German BERT-base (Chan et al., 2020) and (the finetuned) **rott-hc** model, which is the best model from the previous experiments.We also investigate two larger decoder-only LLMs. We selected quantized (legacy) models from the German EM family[11], namely em-LLaMA2[12] and em-Mistral[13]. The sentence embeddings are extracted via the LLaMA.cpp tool[14].

We give the sentence as input to an (L)LM and extract the CLS token embeddings (dim=768 or 4096). Depending on the testing configuration, we add either gaze features (G) or rationales (R) or both, to the sentence embeddings (E). For each gaze feature, we create a feature vector f_i that contains a series of token values for that feature as shown in Figure 1 padded to the maximum token length of the sentences in GAZE4HATE (t=14). The rationales selected in each instance added as bag-of-words vector calculated using the COUNTVECTORIZER module from sklearn (N= 248, number of unique words in the dataset). We have also experimented with token-level rationale representation, see Appendix D.1.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
[11] https://huggingface.co/jphme/em_german_7b_v01
[12] em_german_7b_v01.Q5_0.gguf
[13] TheBloke/em_german_leo_mistral.Q5_0.gguf
[14] https://github.com/ggerganov/

Table 5: Macro and F1 scores for each category of the MLP Classifier. E: word embeddings, G:individual gaze feature, R:Rationale, GPlus: all 6 gaze features (underline : highest score in vertical orientation, bold: highest score among the respective f1-metric (macro, hate or nohate) (horizontal)

	bert-base			bert-ft (rott-hc)			em-LLaMA2			em-Mistral		
condition	**macro_f1**	**hate_f1**	**nohate_f1**	**macro_f1**	**hate_f1**	**nohate_f1**	**macro_f1**	**hate_f1**	**nohate_f1**	**macro_f1**	**hate_f1**	**nohate_f1**
E	0.56	0.54	0.57	0.63	**0.60**	0.65	0.58	0.56	0.59	**0.65**	0.56	**0.73**
EG	0.59	0.57	0.61	**0.69**	**0.68**	0.70	0.60	0.56	0.63	0.68	0.62	**0.75**
ER_{bow}	0.65	**0.63**	0.68	**0.66**	0.61	**0.70**	0.60	0.56	0.64	0.61	0.56	0.65
EGR_{bow}	0.63	0.61	0.65	**0.68**	**0.65**	**0.72**	0.60	0.57	0.62	0.61	0.58	0.65
EGPlus	0.57	0.53	0.61	**0.67**	**0.64**	0.69	0.57	0.56	0.59	0.65	0.58	**0.71**
$EGPlusR_{bow}$	**0.63**	**0.62**	0.64	0.62	0.54	**0.71**	0.58	0.54	0.61	0.61	0.58	0.64

6.2 Results

Table 5 summarizes the performance of various feature combinations on predicting subjective hate (binary classification as hate versus no-hate). We report macro-F1 and F1-scores for both hate and no-hate classes. The first row corresponds to the performance of the model trained on only CLS embeddings (E). CLS&Gaze (EG) row provides the highest score obtained with the inclusion of a gaze feature one at a time. The third row belongs to the CLS&Rationale (ER) model (no gaze feature). The next variation includes rationales added to the EG Model (EGR). Finally, the last two variations include all gaze features (Plus). The contribution of each individual feature is presented in Appendix 7.

For the subjective HSD, the finetuned MEANION models predominantly outperform other MEANION models. The injection of gaze features increases performance: .03 F1-score improvement using the BERT-base, .06 using the **rott-hc**, .02 with em-LLaMA2, and .03 using em-Mistral. The rationales contribute more to the BERT-base MEANION (.09), slightly improve the performance of the MEANIONs with the finetuned (.03) and em-LLaMA2 models (.02), and it drops the performance of the em-Mistral ($-.04$). Except for the BERT-base model, they even hurt the performance up to .07 when combined with gaze features. It should also be highlighted that integrating gaze and rationale features to BERT-base MEANION brings the performance closer to the text-only **rott-hc** MEANION. The results highlight that gaze features provide substantial complementary information for subjective HSD and produce similar effects to fine-tuning on hate speech data.

For E-only models, MEANIONs with only the em-LLaMA2 and em-Mistral embeddings (without fine-tuning) indicate higher performance compared to the BERT-base MEANION. The contribution of gaze and rationales to em-LLaMA2 embeddings seems to be at the similar level. Furthermore, em-Mistral plus gaze embeddings are the best among the em-Mistral variations, and these results are significantly better than em-LLaMA2 performances and BERT-base models. The results demonstrate that EG models outperform all other variations. These also further confirm our conclusion that gaze features provide complementary information for subjective HSD, which is not represented in smaller or large LLMs.

In conclusion, MEANION with the finetuned BERT, especially the gaze-integrated one, outperforms all other variations. E-only em-LLaMA2 and BERT-base models perform on a similar level. Among these variations, E-only em-Mistral achieves higher macro-F1, yet the finetuned (rott-hc) ones show better F1-score for the hate class. The contribution of eye movements on (L)LM only models is consistently observed and statistically proven with our further pairwise model comparisons using the McNemar's test (see Appendix Figure 11).

7 Discussion

Based on the above described experiments we revisit our research questions.

RQ 1: Do gaze features provide robust predictors for subjective hate speech annotations? Yes. According to the analysis of annotators' gaze patterns, 6 out of 13 gaze features differ with respect to the subjective hate categories.

RQ 2: How do gaze features correlate with human and model rationales? InputXGradient method seems to be more aligned with the fixation-based gaze and human rationales, which makes it more suitable explanation method for subjective hate ratings. But the pupil size related parameters are not correlated with model rationales, this might mean that the signal carried by pupil size might be one of the missing components in the HSD models. More systematic analysis on the individual token level among the systematically manipulated con-

ditions, which is beyond the scope of this paper, might provide valuable insight for future directions.

RQ 3: Are gaze features useful for enriching LMs for HSD? Yes. For a MEANION model all six features as well as the human rationale improve performance (compared to using embeddings alone). A further question arises from this conclusion: Do features that correlate badly with model rationales (i.e. carrying complementary information) improve the performance of a model enriched with these features? Figure 5 plots the relationship between subjective hate rating effects, correlation with InputXGradient rationales, and error reduction in MEANION models. It shows that the features badly correlating with the model rationales do not necessarily improve the MEANION models (they do for base (B) but not for the **rott-hc** model (F)).

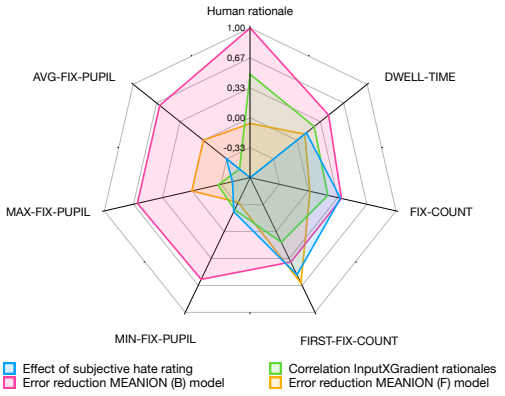


Figure 5: Effect of subjective hate rating, the correlation with model rationales and the error reduction for both the base and rott-hc MEANIONs, for six gaze features and human rationale [15].

8 Conclusion

We introduce a rich dataset of human readings of hate speech. Our GAZE4HATE dataset is enriched with gaze features and subjective hatefulness ratings collected from 43 participants on 90 sentences (3870 unique subjective annotation instances). We compare subjective human hate ratings, human gaze and human rationales with hate speech models rationales. By doing so, we also experiment with various model explanation methods and compare their performance in aligning with human behaviour. The human attention values (represented with a set of gaze features and rationales) are a highly valuable source not only for evaluating the models, but also for training them with cognitively guided attention mechanisms (Ding et al., 2022; Long et al., 2019; Hollenstein et al., 2019). In addition, we also introduce the first gaze-integrated hate speech model (MEANION), which successfully shows the contribution of gaze features on subjective hate speech classification.

Acknowledgements

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 Northrhine Westphalia, Germany.

Additionally, we would like to thank Elisabeth Tiemann and Maria Garcia-Abadillo Velasco for their valuable contribution to the annotation and data collection phases.

Limitations

To evaluate the individual effect of the human gaze and rationale, we implement a basic solution without complex training schemes or multimodal fusion techniques. Our results encourage pursuing more sophisticated implementation for modeling the human gaze for classifying subjective hate speech. Because of space constraints, we could not elaborate on the differences between linguistic manipulations, which can help explain the relations between human gaze, human rationales, and model rationales.

There are linguistic or even non-linguistic factors like (word length, word frequency, expectations etc.) in our experimental set-up that influence cognitive processes. We attempt to minimize these risks with the careful selection of minimal pairs, the random ordering of the sentences, dealing with null values etc.

It should be noted that the decoder-only models are trained on different objectives than BERT-based models. There is a significant amount of ongoing research on how sentence or token embeddings should be extracted or how they could be interpreted. In our paper, we do not aim to address these issues.

Due to the controlled data collection procedure to explore the statistical robustness of different types of gaze features for subjective hate speech detection, the experimental setup may not fully reflect real-world scenarios of hate speech detection. We know that the participant pool lacks diversity, primarily consisting of university students. This might raise concerns about ecological validity. Despite

this limited diversity, our results indicate subjective variation, especially concerning specific statements, as could be seen in Figure 8 and Figure 9 in Appendix B.3. Even in the same apriori category, we observe variation in terms of averaged hatefulness score. Besides, the deviation for each sentence also varies. To address this limitations, future work will address extending the diversity in the participant pool (different backgrounds, cultures, languages, ages etc) and the target groups addressed in the dataset.

Ethics Statement

All recordings have been made after the signed consent of the annotators. Participants' identities are anonymized using pseudo-participant ID. The shared data do not contain any cues to reveal their identities. The dataset contains hateful statements about women and men, which do not reflect the opinion of any of the authors.

Hate speech is widespread in social media and causes a lot of harm to individuals, groups, and societies. Therefore, we consider social media as a possible application area, where models fine-tuned with gaze information can be used for individualized content moderation. Yet, our research does not imply that individual gaze information needs to be shared with/evaluated by social media companies. Eye-tracking technology, already part of many virtual headsets (HTC VIVE[16], Apple Vision[17], etc.), seems to be entering our daily lives through our phones and laptops (e.g., Rathnayake et al. (2023); Brousseau et al. (2020)). From an application point-of-view, incorporating users' gaze into phone applications via offline applications or through federated learning (by deploying a trained model) that can be integrated into social media or messaging APIs might take the privacy concerns into account.

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Özge Alacam, Eugen Ruppert, Ganeshan Malhotra, Chris Biemann, and Sina Zarrieß. 2022. Modeling referential gaze in task-oriented settings of varying referential complexity. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 197–210, Online only. Association for Computational Linguistics.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584.

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.

Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *Hawaii International Conference on System Sciences*.

Braiden Brousseau, Jonathan Rose, and Moshe Eizenman. 2020. Hybrid eye-tracking on a smartphone with cnn feature extraction and an infrared 3d model. *Sensors*, 20(2):543.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[16] https://www.vive.com/nz/support/vive-xr/category_howto/eye-gaze-targeting.html

[17] https://www.apple.com/apple-vision-pro/

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

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.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Karin de Langis and Dongyeop Kang. 2023. A comparative study on textual saliency of styles from eye tracking, annotations, and language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 108–121, Singapore. Association for Computational Linguistics.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Xiao Ding, Bowen Chen, Li Du, Bing Qin, and Ting Liu. 2022. Cogbert: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3210–3225.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. 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.

Sanne Hoeken, Sina Zarriess, and "Ozge Alacam. 2024. Hateful word in context classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):95.

Eileen Kowler. 2011. Eye movements: The past 25 years. *Vision research*, 51(13):1457–1483.

Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019. Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing*, 12(4):900–912.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152, Online. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019*, pages 14–17. ACM.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang Resources & Evaluation*, 55:477–523.

Rasanjalee Rathnayake, Nimantha Madhushan, Ashmini Jeeva, Dhanushika Darshani, Akila Subasinghe, Bhagya Nathali Silva, Lakshitha Wijesingha, and Udaya Wijenayake. 2023. Current trends in human pupil localization: A review. *IEEE Access*, 11.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand, editors. 2021. *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. Association for Computational Linguistics, Duesseldorf, Germany.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *CoRR*, abs/1701.08118.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Lloyd S Shapley. 1953. *A value for n-person games*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos S Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, 16:260–277.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection

on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, pages 1 – 10. Austrian Academy of Sciences, Vienna, Austria.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.

A Appendix

A.1 Instructions for the Annotators

The experimental instructions were given in written format in German. After the instructions, the participants completed 4 familiarization trials. Before starting with the main experiments, we make sure that they do not have any further questions regarding the task. The following text corresponds to the translated instructions:

> During this experimental session, you will be presented with 90 sentences. While some sentences have highly positive sentiments, some of them are hateful. There are also sentences that are neither positive nor hateful. For the current study, we define hate speech as expressions that carry a very negative stance (in terms of their intent). Please always keep this definition in mind and annotate the sentences carefully. One trial consists of (i) reading a sentence, (ii) evaluating its hatefulness, (iii) evaluating your confidence in this decision, and finally, (iv) highlighting the parts of the sentence that contribute to its hateful meaning (if any).
>
> Step-1: Read the sentence freely and press a key when you are done reading.
>
> Step-2: You will be asked to evaluate the sentence on a 1 to 7 Likert scale. Please think thoroughly.
>
> Step-3: You will be asked to evaluate your certainty/confidence while giving this score.
>
> Step-4: In this final step, each word in the sentence is shown in a bounding box. Please click on the words that contribute to your decision. You can have multiple selections. The boxes will be highlighted when you click them or hover them with your mouse during a press. To unselect a box or a series of boxes, you can click on them again. Feel free to try the annotation tool out during the familiarization period.

A.2 Data Availability

In addressing the reproducibility of our study as well as the availability of software and datasets, we provide the following link to our GitHub repository under a CC-BY-NC 4.0 license: `https://gitlab.ub.uni-bielefeld.de/clause/gaze4hate`.

A.3 Appendix: SR Eyelink definitions of gaze features

The description of row features which are directly taken from SR-Eyelink Dataviewer Export (User Manual : Data Viewer 4.3.210 `https://www.sr-research.com/support/`):

- FIXATION: Percentage of all fixations in a trial falling in the current interest area.
- DWELL-TIME_%: Percentage of trial time spent on the current interest area
- MAX-FIX-PUPIL-SIZE: Maximum pupil size among all fixations falling within the interest area
- MIN-FIX-PUPIL-SIZE: Minimum pupil size among all fixations falling within the interest area
- AVERAGE-FIX-PUPIL-SIZE: Pupil size of the current sample averaged across the two eyes.
- RUN_COUNT: Number of times the Interest Area was entered and left (runs).
- REGRESSION_IN (categorical): Whether the current interest area received at least one regression from the later part of the sentence
- REGRESSION_IN_COUNT: Number of times the current interest area was entered from interest areas with higher IA_IDs.
- REGRESSION_OUT (categorical):Whether regression(s) was made from the current interest area to the earlier part of the sentence
- REGRESSION_OUT_COUNT: Number of times the current interest area was exited to a lower IA_ID before an interest area with a higher IA_ID was fixated in the trial.

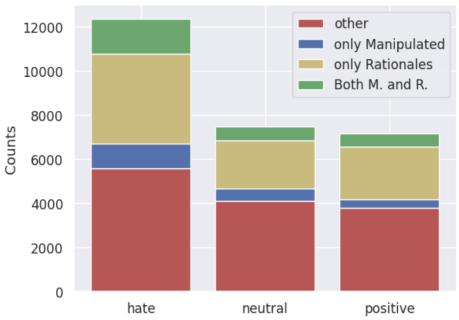


Figure 6: Number of tokens per subjective hate categories

- SKIP (categorical): An interest area is considered skipped (i.e., SKIP = 1) if no fixation occurred in first-pass reading.

In addition to the participant-specific gaze normalization, the data needs to be preprocessed concerning missing values, which are not uncommon in gaze data. For example, if a participant skips a word during reading or a blink is detected, the respective data point is null. If all token values for a gaze feature are null, the trial is removed from the dataset, otherwise, null values are replaced with either zero (if it is skipped) or the average (if a blink is detected).

B Gaze4Haze Annotation Results

B.1 A Closer look at the manipulated tokens and rationales

A Chi-square test has been conducted to see the difference on rationale selections among subjective hate categories. It revealed a significant main effect ($\chi^2(1) = 110.49, p < .001$).

Figure 6 shows the distribution of rationales, manipulated words and other tokens in the entire dataset. Since manipulated tokens occur only in the minimal pair conditions (see 3), their frequency is overall lower compared to rationales and other tokens. The ratio of rationales to all tokens is similar among the subjective hate categories (hate: 32.9%), neutral: 29.1%, positive: 33.49). On the other hand, the ratio of the tokens that are both manipulated and selected is higher in hate category (13.0%) compared to neutral (8.13%) and positive categories (8.33%). A detailed look on the interaction between these two token types are beyond the scope of this paper, here we will provide a glimpse of a bigger analysis.

Manipulated words (parts of minimal word pairs) are the markers that change the hatefulness of the statement. As an example, for the following sentences, "Women belong in the kitchen" and "Pots belong in the kitchen", "women" and "pots" are the minimal pairs, which are manipulated. For the former case, this manipulated token is selected as rationale, in the latter, not.

Since (i) the annotators consistently selected more words in their rationales than only the word we manipulated, and (ii) they select rationales for the positive statements too, the selection of a word for a rationale is not always an indication of hate, but also of general importance for the annotation decision.

We conducted further Anova tests to check whether the gaze features differ on words being manipulated and /or selected for the rationale from the minimal pair conditions. Table 6 shows statistical significance levels of the Anova tests in multiclass and binary comparisons. The gaze measurements on the rationales differ among the subjective hate categories. But when it comes to tokens which are manipulated but not selected (e.g. pots as in the example above), while fixation-based parameters still show significance difference, only pupil size related parameters do not differ, this might tell that pupil size parameters might be more sensitive at the token level while fixation-based parameters are more in line with the overall sentence stance.

Regarding the restricted subset of both manipulated and selected tokens, we also observe cases where gaze measurements show no sensitivity in terms of the hate category (e.g. DWELL-TIME, RUN-COUNT, FIRST-RUN-FIXATION, which differs highly significantly when we look at the all dataset. This means that regardless of their hatefulness, they exhibiting similar gaze patterns. Our manipulations successfully provide fine-grained control conditions, yet their evaluations are beyond the scope of this paper.

	R. Multi (Binary)	M.& R. Multi (Binary)	M. & ~R. Multi (Binary)
FIXATION-COUNT	ns. (0.05)	ns. (ns.)	0.01 (0.01)
DWELL-TIME	0.01 (0.01)	ns. (ns.)	0.01 (0.01)
MAX-FIX-PUPIL-SIZE	0.05 (0.01)	0.05 (0.05)	ns. (ns.)
MIN-FIX-PUPIL-SIZE	0.01 (0.01)	0.01 (0.01)	ns. (0.05)
AVERAGE-FIX-PUPIL-SIZE	0.01 (0.01)	0.01 (0.01)	ns. (ns.)
RUN-COUNT	0.01 (0.01)	ns. (ns.)	0.01 (0.01)
FIRST-RUN-FIXATION-COUNT	0.01 (0.01)	ns. (ns.)	0.05 (0.01)

Table 6: Significance levels of feature-wise comparison of subjective hate ratings on manipulated conditions w.r.t. whether the token is (i) manipulated or not (M) and (ii) selected as rationale or not (R)

B.2 Confidence Ratings

The average confidence score for the a priori categories is above 3.5 out of 5 indicating that the sentences were rated with sufficient confidence rather than random assignment.

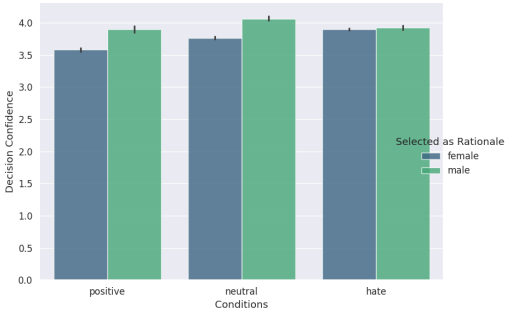


Figure 7: Average confidence scores on the a priori labels (as hate, neutral, positive)

B.3 Subjective Hatefulness Ratings per sentence and condition

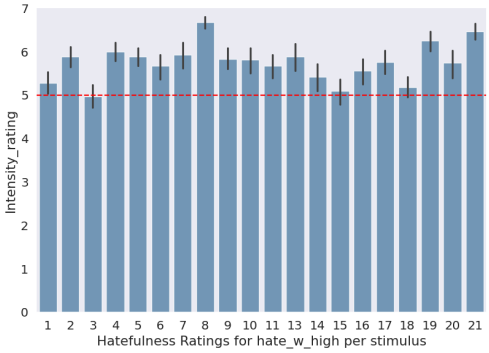
The following Figure 8 illustrates the variation for individual sentences with the a priori labels (x-axis) and their average subjective hate ratings (y-axis), with the confidence intervals representing the variation in the ratings.

We initially created minimal triples (hate, neutral, positive apriori). These are aligned in Figure 8. For example, the first bars in each subfigure are variations of the first hateful sentence (1:hate, 52:neutral, 73:positive), and the second bars in each subfigure belong to the same minimal pair conditions (2, 53, 74) and go like that. The error bars indicate a general agreement between the annotators.

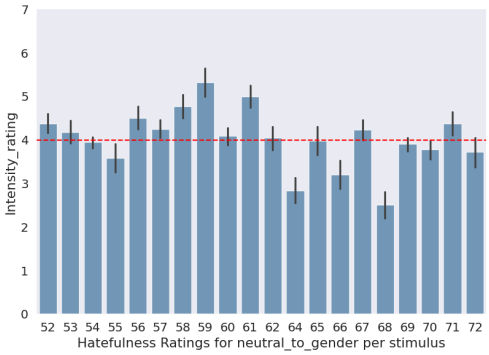
These graphs show that there are several statements which were rated differently and the range of the confidence intervals vary substantially.

In "the high hate against women" (c1) condition (Figure 8a), they are all evaluated as hateful, while the sentences [3: *Frauen haben im statistischen Mittel ein kleineres Allgemeinwissen.*] and [15:*Eine klassische Rollenverteilung in der Ehe ist zu bevorzugen.*] are at the lower band.

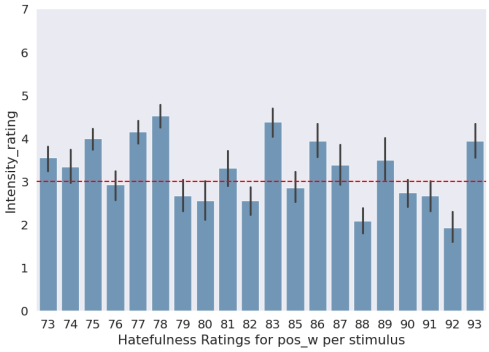
In the neutral (in terms of offensiveness towards gender) statements (Figure 8b), they are overall around the middle band. The sentences [59: *Minions können nichts und sind zu dumm.*] and [61:*Die #keinurlaub-Bewegung ist hysterisch.*] are evaluated as hateful on average, while [64: *Wissenschaftler sind überwiegend angetrieben durch Neugierde.*]and [68: *Gleichstellung der*



(a) High Hate against women (MC) (N=20)



(b) Neutral in terms of gender (MC) (N=20)



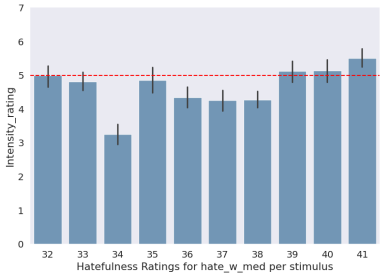
(c) Supportive for women (MC) (N=20)

Figure 8: Subjective hate ratings per experimental condition and stimulus (MC: Manipulated conditions)

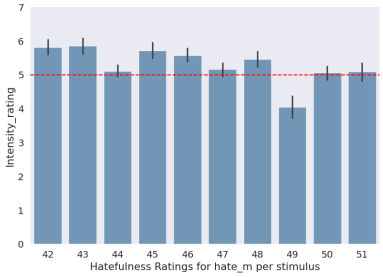
Geschlechter findet zunehmend häufiger statt.] as positive.

In the positive towards women condition (Figure 8c), the evaluation shows mixed, but generally neutral or positive ratings with the exceptions [78: *Frauen sind Männern im Erinnerungsvermögen überlegen.*] [83: *Frauen sollten nicht zu Hause bleiben und sich um ihre Karriere kümmern*].

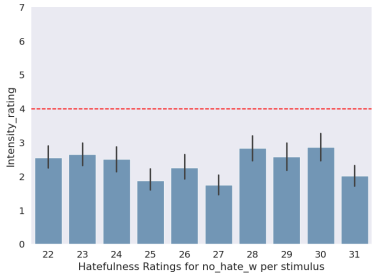
On the other hand, other conditions which are included as control conditions also display interesting tendencies. (Figure 9a) is directly taken from the subset of FEMHATE dataset, namely "medium hate against women". Our participants mostly consider these statements in either mean or neutral



(a) Offensive against women (N=10)



(b) Hate against men (N=10)



(c) No Hate (N=10)

Figure 9: Subjective hate ratings per experimental condition and stimulus

conditions except the sentence [34: Frauenquote muss überall sein.]

The statements in the "Hate against men" condition (Figure 9b) are evaluated as hate on average except the sentence [49: Männer sind bei Stellenvergaben privilegiert.]

The statements in the "No Hate" condition (Figure 9c) are generally evaluated as positive statements.

C HSD Models and Their Rationales

C.1 Details of Huggingface Models

Deepset Deepset Model is finetuned on GermEval18 (coarse and fine) (Wiegand et al., 2019), collected from Twitter data. GermEval18(Coarse) requires a system to classify a tweet into one of two classes: *OFFENSE* if the tweet contains some form of offensive language , and *OTHER* if it does not. For this dataset, similar to our study, the target groups are not explicitly mentioned in the hate speech definition. The author uses the following definition: "In the case of PROFANITY, profane words are used. However, the tweet does not want to insult anyone. In the case of INSULT, unlike PROFANITY,the tweet clearly wants to offend someone. In the case of ABUSE, the tweet does not just insult a person but represents the stronger form of abusive language ascribing a social identity to a person that is judged negatively by a (perceived) majority of society." All these categories were treated in one category in GermEval18 (Coarse) dataset. This model that makes binary classification on broader terms of hate speech aligns with our content as well, yet the inclusion/ratio of gender-related hate in the training data is not known.

Ortiz The model Ortiz is a fine-tuned version of bert-base-german-cased using the HASOC dataset (Mandl et al., 2019) to detect hate speech, specifically in the German language. It has binary class as *hate* versus *no hate*, which aligns with our binary classification. Hate speech is defined as "Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender,social status, health condition or similar." Although gender is not directly mentioned as target group in the hate speech definition, the definition itself looks inclusive. The inclusion/ratio of gender-related hate in the training data is also not known.

ALURU Hate-Speech-CNERG (Aluru et al., 2020), another well-known hate speech model, is fine-tuned on the multilingual BERT model. They use two labels, hate speech and normal, and discard other labels like (offensive, profanity, abusive, insult, etc.). For German, the model is trained on (Ross et al., 2017; Bretschneider and Peters, 2017) datasets. Both German datasets carry hate speech against foreigners. As definition, Ross et al. (2017) dataset uses the Twitter rule as "You may not promote violence against or directly attack or threaten other people on the basis of race,ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability,or disease. We also do not allow accounts whose primary purpose is inciting harm towards others onthe basis of these categories." The Bretschneider and Peters (2017) dataset contains sentences against the government represented by political parties and politicians, the press and media, other identifiable targets, and unknown targets. Yet, gender-related hate speech is

Table 7: Individual contribution of each gaze feature

	BERT-base				finetuned				em-LLaMA2				em-Mistral			
	BG		BGR		BG		BGR		BG		BGR		BG		BGR	
feature	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1	macro_f1	hate_f1
AVERAGE_FIX_PUPIL_SIZE	0.578	0.579	0.614	0.583	**0.689**	0.679	0.621	0.557	0.565	0.552	0.588	0.553	0.628	0.525	0.609	0.577
DWELL_TIME_%	0.548	0.530	0.616	0.585	**0.671**	0.647	0.664	0.606	0.575	0.558	0.571	0.555	0.641	0.554	0.613	0.576
FIRST_FIXATION_DURATION	0.544	0.551	0.631	0.617	0.668	0.640	**0.679**	0.642	0.576	0.578	0.573	0.524	0.670	0.627	0.612	0.580
FIRST_RUN_FIXATION_%	0.540	0.507	0.631	0.611	**0.687**	0.666	0.665	0.619	0.567	0.525	0.583	0.544	0.664	0.600	0.602	0.563
FIXATION_%	0.542	0.515	0.616	0.591	0.642	0.605	**0.643**	0.581	0.587	0.559	0.575	0.546	0.638	0.551	0.615	0.579
MAX_FIX_PUPIL_SIZE	0.536	0.554	0.613	0.589	**0.660**	0.639	0.650	0.587	0.573	0.549	0.573	0.536	0.658	0.629	0.597	0.555
MIN_FIX_PUPIL_SIZE	0.567	0.530	0.605	0.577	**0.685**	0.669	0.662	0.609	0.565	0.543	0.583	0.556	0.634	0.558	0.601	0.564
REGRESSION_IN_COUNT	0.540	0.532	0.595	0.594	0.670	0.646	**0.683**	0.648	0.573	0.555	0.580	0.542	0.639	0.540	0.607	0.583
REGRESSION_OUT	0.519	0.514	0.605	0.585	0.674	0.653	0.649	0.589	0.573	0.529	0.578	0.552	**0.684**	0.619	0.593	0.548
Pupilsize_variation	0.531	0.534	0.629	0.597	0.675	0.647	0.642	0.592	0.573	0.590	0.596	0.568	0.647	0.553	0.588	0.546
Forward_reg_count	0.588	0.570	0.629	0.604	**0.681**	0.668	0.632	0.555	0.590	0.548	0.560	0.504	0.644	0.566	0.597	0.571

still part of the training data represented in other languages. This dataset is different in terms of data collection; they use seed words to scrap data from Facebook; and the collected data has been annotated by two experts as "slightly offensive to offensive", "explicit to substantial offensive statements" and "none of these" conditions. To conclude, this model is trained on datasets with different annotation styles and labels contributing to its diversity.

Rott : It is a fine-tuned model on three datasets: RP (Assenmacher et al., 2021) and DeTox (Demus et al., 2022). The details of the third dataset, which is the Twitter dataset (Glasenbach, 2022) are unfortunately missing in the huggingface model card. It performs a multi-class classification of hate speech. The classes are *No Hate Speech*, *Other Hate Speech* (Threat, Insult, Profanity), *Political Hate Speech*, *Racist Hate Speech* and *Sexist Hate Speech*. For the Assenmacher et al. (2021) dataset, the definitions vary with respect to the type of hate/abusive speech as follows: "(i) Attacks on people based on their gender (identity), often with a focus on women, (ii) Attacks on people based on their origin, ethnicity, nation , (iii) Announcements of the violation of the physical integrity of the victim, (iv) Denigrating, insolvent, or contemptuous statements, (v) Usage of sexually explicit and inappropriate language, (vi) Organisational content, such as requests on why specific posts have been blocked and finally (vii) Comments advertising unrelated services or products. " This dataset does not always include targets in their definition as well. On the other hand, another dataset used in the finetuning of Rott, DETOX has a stricter definition scheme. It distinguishes between toxic comments and hate speech. "Toxicity indicates the potential of a comment to "poison" a conversation. The more it encourages aggressive responses or triggers other participants to leave the conversation, the more toxic the comment is. On the other hand, hate speech is defined as any form of expression that attacks or disparages persons or groups by characteristics attributed to the groups. Discriminatory statements can be aimed at, for example, political attitudes, religious affiliation, or sexual identity of the victims." We subsumed the predictions on our dataset into two as no hate speech versus others (as hate).

ml6 : German DistilBERT model fine-tuned on a combination of five German datasets containing toxicity, profanity, offensive, or hate speech. All labels were subsumed to either toxic or non-toxic. (i) GermEval18 (labels: abuse, profanity, toxicity). (ii) GermEval21 (Labels: toxic or not). The toxic comments contain "Screaming - Implying volume by using all-caps at least twice", "Vulgar language – Use of obscene, foul or boorish language", "Insults – Swear words and derogatory statements", "Sarcasm -Ruthless, biting mockery" and "Discrimination – Disparaging remarks about entire groups with sweeping condemnation", "Discrediting – Attempt to undermine the credibility of persons, groups or ideas, or deny their trustworthiness" and finally "Accusation of lying Insinuation that ideas, plans,actions or policies are dishonest, subterfuge and misleading". The third dataset is Ross et al. (2017) dataset as mentioned above. The fourth one is Bretschneider and Peters (2017) as mentioned above. The final one is the HASOC 2019 (listed above). This dataset also aligns with our binary classification on a wide spectrum. Yet the inclusion/ratio of gender-related hate in the training data is also not known.

To sum up, in the fine-tuning of these existing huggingface models, their authors seem to embrace a variety in hate speech definitions and class labels. The wide range of the spectrum (offensive, abusive, toxic, etc.) utilized in the selected datasets for fine-tuning them also aligns with our wide spectrum. Furthermore, Rott is explicitly fine-tuned on sexism; this also explains its out-of-the-box best per-

formance. Therefore, we continue with this model for further fine-tuning on the HateCheck Dataset and use the Hate-check further fine-tuned version with multimodal integration. The base models are integrated into our model in a plug-and-play fashion, which makes the extension to include other models straightforward.

C.2 Finetuning Details of rott-hc

We finetuned the **rott** model (see Table 3) on the German HateCheck corpus[18] (Röttger et al., 2021). For finetuning, we used 80% for training and 20% as development set (for evaluation over different epochs). We finetuned the model for 3 epochs with a batch size of 8, running just on a Macbook Pro's CPU. Other details: implementation with *pytorch* and *transformers* libraries, AdamW optimizer for training with learning rate of 5e-5 (and all other default hyperfeatures), applying linear scheduler with 0 warmup steps.

C.3 Label Alignment

Table 8 gives an overview of the label aligning of the different model classes and the binary classification schedule that we used for evaluating the different models.

Table 8: Label aligning of model classes and (human) subjective hate ratings with binary classification schedule for evaluation purposes. (*HS = Hate Speech)

Binary	human	deepset	ortiz	aluru	rott	ml6	rott-hc
HATE	hateful	OFFENSE	1	HATE	Other HS* Political HS Racist HS Sexist HS	toxic	hateful
NO HATE	neutral positive	OTHER	0	NON_HATE	No HS	non_toxic	non-hateful

C.4 Model rationales

Table 9 reports mean correlation values of the human rationales and six gaze features with the model rationales extracted with the three different methods.

Table 9: Mean correlation (Pearson's r) between model and human rationales and features. (No correlation values are included for constant feature arrays)

	n	input_x_gradient	saliency	shapley_value
FIXATION-COUNT	*3602*	0,249	0,221	0,035
DWELL-TIME	*3616*	0,257	0,228	0,038
AVERAGE-FIX-PUPIL-SIZE	*3504*	-0,009	-0,004	-0,002
MAX-FIX-PUPIL-SIZE	*3503*	0,079	0,075	0,005
MIN-FIX-PUPIL-SIZE	*3503*	-0,089	-0,078	-0,01
FIRST_RUN_FIXATION-COUNT	*2604*	0,220	0,200	0,031
Human rationale	*3128*	0,335	0,298	0,077

[18] https://huggingface.co/datasets/Paul/hatecheck-german

D MEANION model results

Table 7 shows the contribution of each gaze feature separately for the base and the finetuned models.

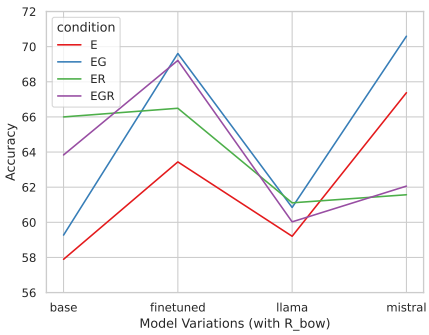


Figure 10: Accuracy scores for all model variations

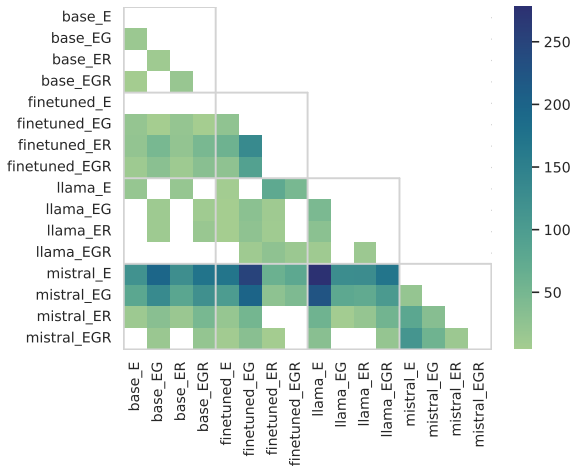


Figure 11: Pairwise Model Comparisons using McNemar's Statistics (only significant differences are visualized, the color denotes the chi-squared value. The darker value means higher Chi-squared value, meaning a bigger significant difference.)

D.1 Position-based and BOW Rationale Representation

Figure 12 illustrates the effect of different rationale representations combined with various LM and gaze embeddings on the HSD classification.As seen from the graph, for the BERT-based models, adding rationales as bag-of-words representation results in higher performance, while for LLMs, we observe the opposite trend, this might indicate that semantic information regarding those words selected as rationales were already represented by the CLS embedding, highlighting the position of the rationales in combination with gaze information bring forth more complementary information.

D.2 Implicit versus Explicit Hate Speech

Insights into performance values of the different models with respect to implicitness (Table 10) show

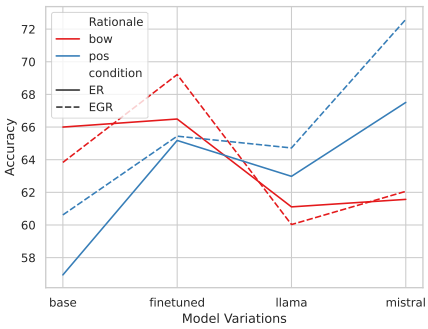


Figure 12: Rationale as Bow versus Row]

		n	deepset	ortiz	aluru	rott	ml6	rott-hc
HATE	**explicit**	944	0.53	0.08	0.00	0.61	0.21	0.68
	implicit	763	0.48	0.00	0.00	0.57	0.10	0.65
NO HATE	**explicit**	1031	0.68	0.69	0.69	0.61	0.71	0.63
	implicit	878	0.71	0.70	0.70	0.62	0.71	0.76

Table 10: Model performance w.r.t. linguistic types.

that for the instances rated as hateful, the models perform better on the sentences where hatefulness is based on lexical cues (F1-score of 0.68 for rott-hc) rather than on implicit knowledge (F1-score of 0.65 for rott-hc). For the instances rated as non-hateful, it seems to be the other way around (F1-score of 0.76 for implicit, 0.63 for explicit cues).

We further plotted the accuracy scores in Figure 13 (i) to understand the models' capabilities to detect explicit and implicit hate speech and (ii) to explore the effect of gaze and rationales on this distinction. Among the base models (BERT, em-LLaMA2 and em-Mistral), the performance difference between hate (red lines) and no-hate (blue lines) classes with BERT and Mistral-based models are pretty clear. Overall patterns indicates the implicit no hate is the easier to classify, while implicit hate is the most challenging case as expected.

D.3 Training Parameters of MLP Classifier

For each LLM model and feature configuration, we conducted grid search using sklearn. Later, each configuration is trained with its best hyperparameters (Table 11.

```
parameter_space = {
'hidden_layer_sizes': [(64, 32),
(128, 64),
(128, 64, 32),
(256, 100),
(256, 100, 32)],
'activation': ['tanh', 'relu'],
'solver': ['sgd', 'adam'],
'alpha': [0.0001, 0.0005,
0.001, 0.005, 0.01], # ,
'learning_rate': ['constant',
'adaptive'],
```

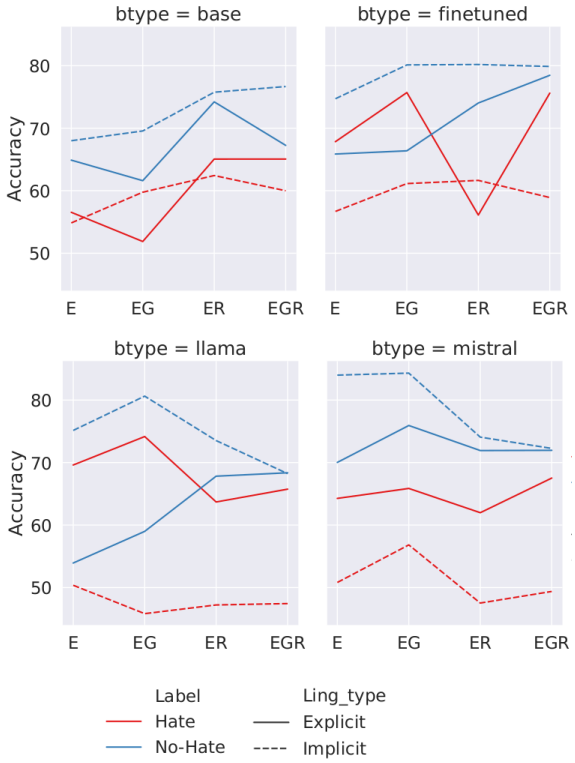


Figure 13: Accuracy Scores of all model variations on Implicit versus Explicit Statements

Table 11: Best hyper-parameters after grid search for each configuration

	features	lr	hidden layer sizes
BERT-base and finetuned-BERT			
bow	B	0.001	(256, 100)
	BG	0.0001	(128, 64, 32)
	BR	0.001	(128, 64, 32)
	BGR	0.0001	(128, 64, 32)
pos	B	0.001	(256, 100)
	BG	0.0001	(128, 64, 32)
	BR	0.0001	(128, 64, 32)
	BGR	0.0001	(128, 64)
em-LLaMA2 and em-Mistral			
bow	B	0.001	(256, 100)
	BG	0.001	(256, 100)
	BR	0.0001	(64, 32)
	BGR	0.0001	(64, 32)
pos	B	0.001	(128, 64)
	BG	0.001	(256, 100)
	BR	0.0001	(64, 32)
	BGR	0.0001	(64, 32)