Disentangling Subjectivity and Uncertainty for Hate Speech Annotation and Modeling using Gaze

Özge Alaçam[1] Sanne Hoeken[1] Andreas Säuberli[2,3] Hannes Gröner[1]
Diego Frassinelli[2,3] Sina Zarrieß[1] Barbara Plank[2,3]

[1]Bielefeld University, Germany,
[2]LMU Munich, Germany, [3]Munich Center for Machine Learning, Germany
{oezge.alacam, sanne.hoeken, hgroener, sina.zarriess}@uni-bielefeld.de
{andreas.saeuberli, diego.frassinelli, b.plank}@lmu.de

Abstract

Variation is inherent in opinion-based annotation tasks like sentiment or hate speech analysis. It does not only arise from errors, fatigue, or sentence ambiguity but also, for example, from genuine differences in opinion shaped by background, experience, and culture. In this paper, first, we show how annotators' confidence ratings can be of great use for disentangling subjective variation from uncertainty, without relying on specific features present in the data (text, gaze etc.). Our goal is to establish distinctive dimensions of variation which are often not clearly separated in existing work on modeling annotator variation. We illustrate our approach through a hate speech detection task, demonstrating that models are affected differently by instances of uncertainty and subjectivity. In addition, we show that human gaze patterns offer valuable indicators of subjective evaluation and uncertainty.

Disclaimer: This paper contains sentences that may be offensive.

1 Introduction

Many areas of NLP rely on human-annotated data and treat these annotations as so-called ground truth for training, fine-tuning, or testing models. In established annotation workflows, ground-truth data is often generated by multiple annotators and very commonly contains variation, e.g., with different annotators assigning different labels to the same sentence, as exemplified in Figure 1. How to deal with this variation has been a long-standing and notorious question for research in NLP (Alm, 2011; Poesio and Artstein, 2005; de Marneffe et al., 2012; Aroyo and Welty, 2015).

Traditional annotation approaches typically aim to reduce variation as much as possible by tailoring annotation guidelines, removing annotation errors, or employing majority voting, etc. Such efforts often take a prescriptive annotation approach (Röttger et al., 2022), unlike a descriptive one. As extensively discussed in Anand et al. (2024); Arazo et al. (2019); Mokhberian et al. (2022), variation in annotation is not only due to annotation errors (such as careless drops or annotator fatigue) or uncertainty (from lack of knowledge, ambiguous instructions, or ambiguous instances), but also to differences in opinions between annotators. Recent approaches, therefore, have argued that it may not be possible or even desirable to exclude variation, arguing for descriptive annotation, and made proposals to embrace so-called human label variation (Plank, 2022), especially in highly subjective tasks (Leonardelli et al., 2021; Uma et al., 2021; Basile et al., 2021; Casola et al., 2023).

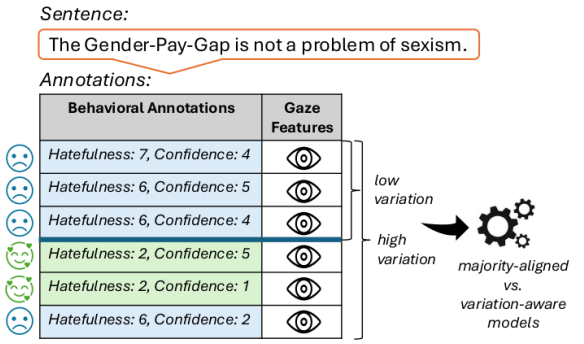

Figure 1: Different kinds of variation in data. The first three rows align with the majority vote, representing collective opinion. The fourth and fifth rows reflect minority votes with varying levels of reported certainty, distinguishing between subjective opinion and subjective uncertainty. The final row aligns with the majority vote, albeit with reduced certainty, representing collective uncertainty.

In this study, we examine different sources of human label variation in a hate speech detection task and explore *whether it is possible to disentangle uncertainty from genuine subjectivity*. Hate speech is inherently multifaceted and hard to define with discrete categories (e.g., offensiveness vs. hatefulness), contributing to annotator uncertainty. Our approach systematically categorizes the instances

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28696–28713
November 4-9, 2025

of subjectivity and uncertainty (see Figure 2) using self-reported confidence and hatefulness ratings. Next, we demonstrate how this approach can serve as a diagnostic tool for analyzing model behavior across different types of instances in hate speech detection (HSD). Finally, we turn to gaze-infused hate speech detection which has recently been proposed as an approach that "naturally" handles annotators' subjectivity (Alacam et al., 2024). Using our quadrant-based approach, we examine how gaze infusion affects the behavior of a hate speech detection model for different sources of human label variation. Disentangling these notions within the data offers several key benefits across various dimensions: (i) it facilitates a deeper understanding of the data characteristics, such as quantifying the proportions of collective opinions and subjective variation; (ii) it enables the analysis and comparison of models' sensitivities to different types of instances, including their robustness to noise, their capabilities to represent individual variation and uncertainties.

2 Related Work

Notions of label variation. The concepts of subjectivity and disagreements are integral to opinion-mining tasks where it is clear that annotators can exhibit different/subjective opinions leading to individual variations (Sap et al., 2022).

Despite ongoing efforts in the NLP and ML communities, there remains little consensus on the definition and measurement of human label (annotation) variation. A non-exhaustive list of the most occurring terms in this domain includes *error*, *noise*, *bias*, *random & systematic variation*, *subjectivity*, *disagreement*, *uncertainty* etc. Whereas some of them are used interchangeably, many have nuanced differences.

In particular, while there have been recent attempts to highlight the importance of retaining disagreements originating from genuine subjective opinion (Fleisig et al., 2023), remaining variations are usually treated under the umbrella term *noise*, which may encompass different notions: uncertainty, experimental errors with response latency (too fast or too slow), annotator fatigue, a lack of knowledge (Sandri et al., 2023; Basile et al., 2021; Zhang and de Marneffe, 2021). Among these, the notion of uncertainty (e.g. treating the instances with uncertainties as noise (Jinadu and Ding, 2024) has received particular attention. It is critical to note that apart from a few exceptions (Baan et al., 2023; Peterson et al., 2019), the term *uncertainty* primarily is attributed to the model's uncertainty in predicting correct class rather than the human uncertainty reported during the annotation process.

A deeper understanding of the nature of these closely related but distinct concepts, and their careful disentanglement through behavioral and psychological analysis (e.g., identifying which instances are genuinely noise versus meaningful data reflecting subjective variation), is crucial for accurately modeling opinion-mining tasks.

Methods for dealing with variation. Many existing ML models implicitly assume statistical independence between instances. Mitigating this simplification requires a more deliberate approach such as removing variation from the data (Zampieri et al., 2019), training individual models, adjusting loss functions (Jinadu and Ding, 2024; Anand et al., 2024; Simchoni and Rosset, 2023; Arazo et al., 2019), or employing multi-task learning (Mostafazadeh Davani et al., 2022).

Variation-removing strategies such as discarding data points or majority voting are commonly used to create more streamlined datasets. However, as mentioned before, opinion-mining tasks are inherently subjective, and such normalization processes may significantly diminish this valuable aspect. Another strategy to mitigate variation is noise correction through loss modeling approaches (Jinadu and Ding, 2024; Swayamdipta et al., 2020; Mokhberian et al., 2022). Such automated data evaluation strategies seem to be leading to improved model performance on the majority voting prediction. To increase the subjectivity, there are also successful attempts to utilize annotator-dependent features e.g. annotator embeddings (Hoeken et al., 2025, 2024; Vitsakis et al., 2024; Deng et al., 2023; Casola et al., 2023), where the idea is to adjust the model to be either less or more confirming with majority opinions, where the former is about improving the alignment with disagreement patterns.

Recent research has also explored using LLMs to address individual variation in opinion-mining tasks: evaluating the capabilities of LLM as judge (Lu et al., 2025), estimating disagreement through fine-tuning with preference optimization (Loftus et al., 2025), leveraging annotator-specific prompts (Orlikowski et al., 2025), and applying in-context learning with entropy derived from annotator disagreement (Caselli and Plaza-del Arco, 2025).

However, the results suggest that there is still considerable room for improvement in making LLMs more sensitive to subjectivity.

Although we acknowledge the value of model-driven correction strategies aimed at mitigating the negative impact of noise to build more robust classifiers, our approach takes a different direction. We propose a diagnostic concept designed to reveal model sensitivities across fine-grained categories of subjective evaluation. By shifting the focus to the data collection process, we highlight the importance of incorporating an often-overlooked yet critical parameter – confidence ratings (on a continuous scale) – to gain deeper insights into the data and the model's behavior.

Gaze signals for opinion-mining tasks. Previous research has shown that human gaze provides valuable insights into word complexity, implicit language and sentiment perception (Mishra et al., 2016). As noted in Alacam et al. (2024), eye-movement parameters capture distinct aspects of hate speech: pupil size reflects sentiment intensity, while fixation-based parameters help distinguish hate from non-hate speech. Moreover, gaze features are user-specific, revealing subconscious biases and cognitive processes without the need for explicit judgments, and they offer continuous-scaled data instead of being confined to binary (yes/no) responses.

Another study by Cala et al. (2023) tested whether users' gaze can be leveraged to estimate implicit attitudes toward climate change using a standardized Implicit Association Test (IAT) that measures concept associations underlying implicit biases and prejudices. The assumption is that decision-making reaction times are faster for congruent associations than for incongruent ones. Their results showed that three of 13 selected gaze features differed significantly and proved useful for predicting users' implicit attitudes. Similarly, Hansen et al. (2015) investigated whether implicit or explicit racial prejudice can be revealed by eye gaze through scanpath analysis. Their results indicate that, regardless of prejudice type, participants with high and low levels of racial prejudice examined faces differently – providing further evidence of gaze as a predictor of subjective evaluations. Their fine-grained quadrant-based analysis, incorporating scores from standardized implicit and explicit association tests, also demonstrated gaze's sensitivity in revealing different attitudes. The association between gaze and decision making processes has also long been established in prior research (Thomas et al., 2019; Krajbich et al., 2010; Shimojo et al., 2003). These studies consistently showed that gaze patterns are strongly linked to individuals' decision making processes, and can be leveraged both to predict individual differences and to explain variability in choice behavior.

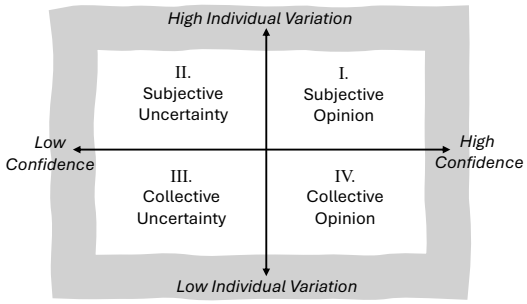

Figure 2: Conceptualization to disentangle subjective variation from uncertainty based on annotators' confidence scores and individual variation at the instance level.

Beyond sentiment and implicit attitudes, gaze has also been shown to predict readers' subjective understanding (the readers' perception of their own understanding) (Lima Sanches et al., 2018). Results revealed that the texts rated with high and low subjective understanding elicited different gaze patterns in terms of saccade direction and length, number of fixations, and other parameters. Yet, the authors note that the benefits of incorporating gaze diminish under certain reading behaviors (e.g., fast reading without engagement), emphasizing both the complexity of the task and the complementary role of gaze.

On the computational modeling side, recent advancements in gaze-integrated models for opinion mining tasks further highlight the utility of gaze signals (Alacam et al., 2024; Wang et al., 2024; Deng et al., 2024; Yang and Hollenstein, 2023). The main principle behind these models is to detect the most attended tokens in a sentence using eye-tracking and guide the model's attention with this rich information (Wang et al., 2024; Deng et al., 2024). These developments raise the question of whether gaze signals also contribute to effectively predict subjectivity and uncertainty in opinion tasks.

3 Our Approach

In this paper, we conceptualize different sources of variation in labeling data by proposing a systematic approach to disentangling subjective variation

from uncertainty. The main sources of information are: (i) individual *labels* (or ratings) by annotators, which are used to calculate the individual variation from all the other annotators, (ii) scores of the annotators' individual uncertainty (*confidence score*) and (iii) user's gaze as optional behavioral data.

Our conceptualization – illustrated in Figure 2 – is built upon two of these core dimensions: *confidence* of an individual annotator and *individual variation* in evaluations (e.g. hatefulness rating) between different annotators (both collected at the instance level on a continuous scale). We test our conceptualization on subjective hate speech detection using the GAZE4HATE dataset (as detailed in the next section). Utilizing the hatefulness scores, the individual variation is calculated for each instance i based on how much the individual's (p) hatefulness score HS_i^p deviates from the average hatefulness score assigned by all other annotators ($q \in [1, .., N]$ where $q \neq p$) on the same instance using the following formula, where N is the total number of annotators who evaluated the sentence (excluding the same annotator):

$$\Delta\mathrm{HS}_i^p = \left| \frac{\sum_{q=1}^{N-1} \mathrm{HS}_i^q}{N-1} - \mathrm{HS}_i^p \right|$$

Using these two dimensions, we categorize instances into four groups: *collective opinion*, *subjective opinion*, *collective uncertainty* and *subjective uncertainty* (Figure 2), as described next. Overall, the right side of the Cartesian plane denotes high confidence (certainty) instances.

The bottom-right quadrant (**Q-IV: Collective Opinion**) includes instances where annotators report high confidence and exhibit low variation from the mean majority rating. These represent relatively clear-cut cases with broad consensus and are expected to be easier for a HSD model to classify. As an example, the sentence *Women are underperformers in our society* was predominantly evaluated with low variation and high confidence, indicating a collective opinion.

The top-right quadrant (**Q-I: Subjective Opinion**) is of high certainty with disagreement and more closely associated with subjective opinion, as those annotators confidently deviate from the average hatefulness score of this sentence. For example, the statement *Women are psychologically not made for complex work* was equally observed in both Q-I and Q-II, receiving high confidence scores but varying levels of agreement. This suggests that perceptions of its hatefulness differ across annotators. In comparison, the implicit and positive statement *Equality is a human rights issue* exhibited greater variation in responses while maintaining a high level of confidence.

The left of the Cartesian plane represents *uncertain* cases. Although our primary focus is on disentangling subjective variation from uncertainty, it remains important to distinguish between **Subjective Uncertainty (Q-II)** and **Collective Uncertainty (Q-III)** for practical reasons. The bottom-left quadrant (*collective uncertainty*) includes instances where annotators report low confidence but still align with the majority, indicating some degree of sentence-level systematicity. For example, an expert-provided (*a priori*) neutral statement such as *Women have significantly higher pain tolerance* was assessed with low confidence but high agreement, reflecting collective uncertainty. Conversely, the *a priori* hateful statement *Women have a significantly lower will to assert themselves* appeared equally in Q-III and Q-IV. While it exhibited low variation in terms of hatefulness, it shows high variation in the confidence dimension.

In contrast, instances in **Q-II: Subjective Uncertainty** represent cases where annotators are uncertain and do not align with the majority, potentially including random guessing by annotators. For example, the statement *The Gender-Pay-Gap is not a problem of sexism* was more frequently observed in Q-II, with evaluations characterized by low confidence and high variation.

3.1 Data

Our approach to distinguish subjective variation from uncertainty requires fine-grained instance-level hatefulness and confidence ratings from each annotator for each sentence, rather than coarse binary labels alone. Unfortunately, many datasets provide this information at a coarse-grained level only. Therefore, to categorize variation regarding hate speech, we utilize the GAZE4HATE dataset (Alacam et al., 2024)[1] which – to the best of our knowledge – is the only dataset providing confidence and gaze on subjective data. It provides confidence and hatefulness scores (on continuous scale) for 90 sentences of positive, neutral and hateful statements towards women (in German), each rated by 43 participants resulting in 3,630 instances. It contains further data (response time, and

[1] GAZE4HATE dataset: https://osf.io/fgdjw

various gaze parameters) collected at the instance-level. The data used in this paper is coming from three experimental phases, namely sentence reading, hatefulness rating, confidence rating. From the first phase, we utilize the gaze recordings collected during sentence reading phase. From the second and third phases, we obtain hatefulness and confidence ratings on each instance. The hatefulness ratings are collected for each instance on a 1-to-7 Likert Scale.[2] Confidence score ratings are self-reported on a 1-to-5 Likert scale,[3] collected right after annotators read a sentence and rate its hatefulness. The individual variation is calculated as explained above by complementing instance-level rating with that of all annotators. All numerical values were normalized at the annotator level using z-score transformation.

A key feature of GAZE4HATE is the wide variation in subjective hatefulness scores (as clearly illustrated in Appendix Figure 5), with some sentences rated differently than their *a priori* category and with high disagreement levels. The authors treat all deviations from the majority as variation, without distinguishing uncertainty from genuine opinion differences. In contrast, we analyze the data to explore whether certain parameters can help disentangle subjectivity from uncertainty – an essential yet challenging task in opinion mining.

3.2 Operationalization

Following our approach, instances are projected onto the Cartesian plane based on two variables: confidence rating, and individual variation at the instance level, as shown in Figure 3. The number of instances per quadrant from highest to lowest is as follows: Q-IV ($N = 1137$), followed by Q-I ($N = 881$), Q-III ($N = 775$) and finally Q-II ($N = 570$). The distributions of various parameters for each quadrant are detailed in Appendix Figure 9. The code for the disentanglement approach and analysis is available at https://github.com/oalacam/disentangle_subjectivity.

For visualization, we color-coded multiclass HSD labels by aggregating ordinal hatefulness ratings into three categories: positive (1 to 3), neutral (4), and hateful (5 to 7). For the classification task in the next section, we combined the positive and neutral classes into a single *nohate* category.

[2] 1: very positive, 2: positive, 3: somehow positive, 4: neutral, 5: mean, 6: hateful, 7: extremely hateful

[3] 1: not certain, 2: somewhat certain, 3: moderate, 4: certain, 5: very certain

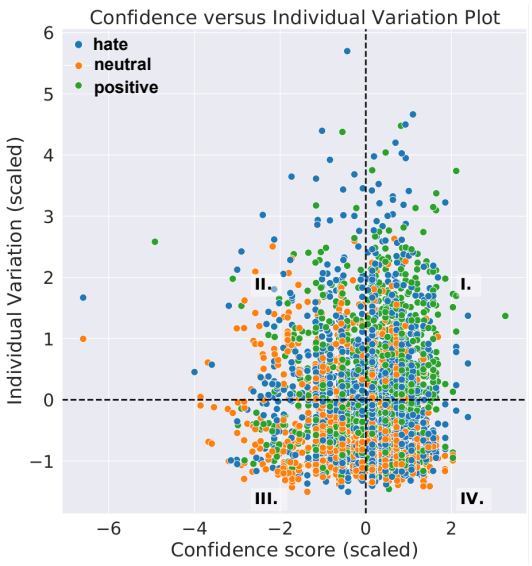

Figure 3: z-score normalized confidence ratings vs. individual variation.

While we focus on variation in hatefulness, further analysis of annotator tendencies, confidence score variability, and the distribution of quadrant instances by annotator gender is provided in the Appendix A.2.

4 Disentanglement Approach as a Diagnostic Tool for HSD Models

We train BERT-based hate speech classifiers on the GAZE4HATE dataset and use the quadrant-based approach from Section 3 to evaluate model performance across different types of variation. This highlights how our method can serve as a diagnostic tool to better understand model behavior, with a focus on the following research question:

RQ1 *Do the models' behaviors change across different quadrants?*

Models. We experiment with two pretrained BERT-based transformer model variations. The first one is the pretrained German BERT model[4] from Hugging Face. Furthermore, we also test with a task-specific fine-tuned model provided alongside the dataset, namely **rott** model[5] fine-tuned on the German HateCheck dataset[6] (Röttger et al., 2021).

For the first part of the analysis, we employ these pretrained models and apply light additional fine-tuning on the GAZE4HATE data using only sentences and their labels. This results in two model variations: *bert-text* and *hsd-text*. For the second

[4] https://huggingface.co/dbmdz/bert-base-german-uncased

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

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

part of the analysis, we infuse gaze features during the fine-tuning of the transformer models as proposed by Wang et al. (2024) using the Gaze-infused BERT model. This results in two more models: *bert-gaze* and *hsd-gaze*. For both text-only and gaze-infused finetuning, we employ the same implementation by disabling or enabling the gaze infusion. The model parameters are optimized using grid-search, with details provided in Appendix A.8.

We train individualized models for each annotator, ensuring that the same sentence cannot occur in different splits. For each annotator, we employ a quadrant-driven splitting method by randomly sampling 80-10-10 percentages from each quadrant for train, val, and test sets. This controls that each quadrant has a proportional sample size from each hate category. Then, we obtain final splits by combining the respective sets of all quadrants. To see the effect of each quadrant, we filtered the annotators ($N = 10$) who did not have enough samples in any of these quadrants. Since we keep the proportion of each quadrant per annotator while creating train-val-test splits, the individualized models still align with the annotator's tendency, without introducing additional class imbalance noise as noted by Mostafazadeh Davani et al. (2022).

4.1 Subjective Labels vs. Model Confidence

As discussed in Section 2 (Jinadu and Ding, 2024; Anand et al., 2024), *uncertainty* usually refers to model uncertainty in predicting the correct class and is one of the heavily used parameters in existing research to handle subjective variation. In this study, we extend this analysis to quantify the relation between subjective labels and class probabilities per quadrant by using point-biserial correlation[7] (Figure 4). These results are obtained from the models trained without any quadrant-based weight reduction (that will be discussed in the next Section). The confidence scores of the *bert-text* model exhibit a significant correlation with subjective gold labels, particularly in Q-IV (*Collective Opinion*) and Q-III (*Collective Uncertainty*), where there is little individual variation. Notably, an interesting difference emerges regarding Q-I instances (*Subjective Opinion*): the *bert-text* exhibits a significantly higher correlation with the subjective labels compared to the HSD pretrained model. While HSD pretraining enhances the alignment with the collective opinion, it hurts the alignment with the high individual variation instances (Q-I, Q-II), without showing significant correlation with Q-II. This finding confirms that the model behavior varies per quadrant, highlighting their distinct nature (addressing **RQ1**).

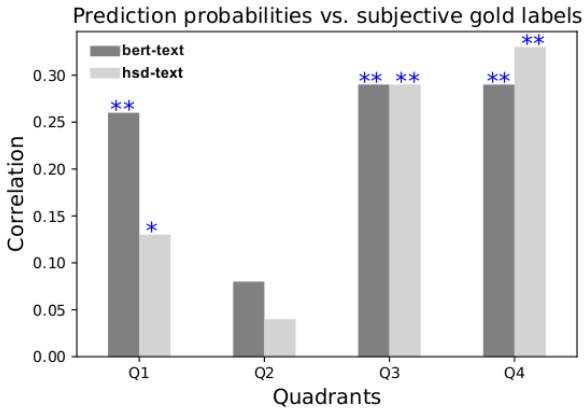

Figure 4: The correlation between the class probabilities of model predictions and the gold labels per quadrant.

4.2 Quadrant-based Weighing of Training Instances

We expect the instances in the four quadrants to affect the training of hate speech detection models in different ways, since these models, trained with text-only data, are not exposed to any information reflecting user subjectivity. We examine the change in the models behavior by reducing the weight of the instances per quadrant in the loss function during training HSD models. The purpose of weight reduction is not to optimize hate speech classification but rather to serve as a diagnostic tool for understanding how different instances influence the model's behavior.

We control the weight of each instance in the training data using the weight parameter of the BCELoss function using two reduction settings: *fixed reduction* versus *distance-based reduction*. In the *fixed reduction* setting, the weights for the instances in the respective quadrant are assigned to a fixed value (0.3)[8] that diminishes the contribution of instance by 70%. The rest of the instances' weights are set to 1. In the *distance-based* setting, the weight is conditioned on the Euclidean distance of normalized scores of the instance to the Cartesian plane centroid (0,0) – the further apart the instance is, the higher the discount of the instance. The distance-based method reduces the impact of instances at the extremes but still retains the impact

[7] scipy.stats.pointbiserialr

[8] determined by manual inspection on the projected space

of the ones closer to the centroid. The three possible model behaviors in response to the reduction of a specific quadrant are provided below.

B1. If performance improves when the influence of specific instances is reduced, it suggests that these instances were a source of confusion for the model in the subjective HSD task.

B2. If performance declines, it indicates that these instances were crucial for the model.

B3. If performance remains unchanged, the model appears invariant to those instances.

To ensure that the observed differences originate from the characteristics of the quadrants rather than the quantity of downgraded data, we conduct additional tests by randomly reducing samples irrespective of their quadrant assignment (including Q-IV). As detailed in Appendix A.9, performance not only remains stable or declines, but in some cases improves, despite a substantial number of instances being downgraded. This indicates that the performance drops are not solely attributable to the quantity of data being manipulated.

Table 1 shows average F1 scores for the sample weighing conditions on two model variations. We also report the performances without any sampling weight reduction during training (*equal weights*) as comparison. We provide the averaged performance metrics over all participants per model.

	Equal	***Fixed***		
Model	**All**	**Q-I**	**Q-II**	**Q-III**
bert-text	.574	.544↓	0.556–	0.592–
hsd-text	**.679**	**.675**–	0.671–	.649↓↓
	Equal	***Distance-based***		
Model	**All**	**Q-I**	**Q-II**	**Q-III**
bert-text	.574	**.628**↑↑	.586–	.603↑
hsd-text	**.679**	.668 –	.618↓↓	**.675**–

Table 1: Average F1 scores with and without HSD fine-tuning. Lowest value: underlined. Highest value: bold. $\Delta <= 0.02\%$ indicates negligible or no change (–). ↓ or ↑ signifies $\Delta > 0.02\%$. ↓↓ or ↑↑ indicates a higher difference, $\Delta > 0.05\%$. Δ is the absolute difference between the respective score and performance of the same model trained with equal weights.

Although our approach serves as a diagnostic tool to explore individual model behavior, we still expect differences on the HSD task originated from the pretraining model. The results reveal that with the de-facto training regime (*equal weights*), the models with HSD pretraining (*hsd-text*) (precision: 0.59, recall: 0.85, F1: 0.68 on average) outperform the *bert-text* model (precision: 0.59, recall: 0.61, F1: 0.57) on average.

Fixed Reduction. To assess the impact of each quadrant, we performed within-model comparisons by measuring the Δ between scores with equal and reduced weights. Q-I includes instances with subjective variation, and down-weighting them during *bert-text* training degrades performance on subjective hate classification (**B2**), indicating that for a model, which is blind to notions of subjective variations, these instances were informative. However, the *hsd-text* model is unaffected by the reduction of Q-I instances, possibly due to its exposure to hate speech concepts that capture some variation. For both *bert-text* and *hsd-text*, reducing Q-II instances – despite their high variation and uncertainty – does not impact performance compared to training with equal weights (**B3**), suggesting the models may already handle noisy data effectively. While reducing Q-III instances has little effect on *bert-text* (**B3**), the *hsd-text* model shows a performance drop, indicating these instances are important for its learning (**B2**).

Distance-based Reduction. This method primarily penalizes the extremes in the respective quadrants. Q-I extremes are source of confusion for the *bert-text* model (**B1**), while reducing their impact has no effect on the *hsd-text* model (**B3**). Consistent with the fixed-reduction method, reducing Q-II does not affect *bert-text*'s performance (**B3**); however, these instances appear critical for *hsd-text* (**B2**), and suppressing them nearly cancels out the benefits of task-specific fine-tuning. Additionally, *bert-text* and *hsd-text* respond differently to reductions in Q-III instances (**B1 and B3**, respectively). Figure 3 illustrates a possible reason for this pattern. Dispersion in Q-III and Q-IV (low variation) is lower than in Q-I and Q-II (high variation), with Q-I (subjective variation) showing the highest spread. This helps explain why Q-I and Q-II are more affected by distance-based reduction.[9]

Differences in model outcomes under both reduction methods indicate that instances across quadrants – each capturing distinct characteristics based on behavioral notions of subjective variation – carry varying degrees of informativeness for these two models (addressing RQ1 for text-only models).

[9] See Figure 9 for mean, max, and standard deviation of the normalized distance measures.

5 Gaze for Subjectivity and Uncertainty

As discussed in Section 2, human gaze inherently reflects an individual's internal processes, making it instrumental for understanding subjective evaluations and uncertainties. In order to understand to what extent human gaze correlates with the core components of our approach, we first conduct a statistical analysis and identify key gaze features for modeling hatefulness, confidence ratings, and individual variation (Section 5.1). We then integrate the most informative gaze correlates into the classification model to assess how it affects model behavior across different variation types represented in each quadrant (Section 5.2).

5.1 Statistical Analysis on Gaze Signals

Subjective Hatefulness. To explore the gaze correlates of subjective hate, we fit a mixed effect logistic regression model with one independent (gaze) feature at a time including random effects (as annotator and sentence). The list of the independent features and the detailed statistical analysis are presented in Appendix A.3 and A.4 respectively. Later, we fit an incremental model, adding only significant predictors – ranked by their estimated magnitude to the final model. Here, we only report the significant gaze features. The strongest gaze predictor of hatefulness category is the *total fixation count*, ($\chi^2(1) = 39.66,\ p < .0001$). As the (subjective) hatefulness of the statement increases, the number of fixation counts increases. *Minimum pupil size* also has a significant effect ($\chi^2(1) = 10.34,\ p < .005$) with a negative relation.

Confidence Ratings. To explore the behavioral correlates of the confidence rating, we fit "Cumulative Link Models for Ordinal Regression" and check the individual effect of each predictor followed by building an incremental model based on the magnitude of the estimates. The most contributing gaze predictor is the *mean fixation count* ($\chi^2(1) = 5.89,\ p < .05$), improving the model fit (with negative correlation). A higher fixation count is observed as the confidence score drops. Moreover, adding *minimum pupil size* ($\chi^2(1) = 6.39,\ p < .05$), *variation in pupil size* ($\chi^2(1) = 5.91,\ p < .05$) significantly improved the fit (with positive correlation). Finally *total regression-out* count had a significant effect as well ($\chi^2(1) = 7.07,\ p < .01$), more back-and-forth movement is observed as the confidence score drops. These results give us a set of parameters we collect from the reading period as indicators of the user's uncertainty about the sentence.

Individual Variation. To check the gaze correlates of individual variation, we used a mixed-effect linear model. The results indicated none of the gaze features collected during reading are significant predictors of individual variation. We intuitively do not expect annotators to be consciously aware of how much their responses deviate from others while they are reading. This high-level awareness is unlikely during such tasks. Nevertheless, we should note that individual variation is computed based on hatefulness scores, which themselves are significantly correlated with certain gaze features.

Interim Discussion. Several reading-phase gaze features[10] strongly reflect subjective hatefulness and confidence judgments. Here we summarize the list of gaze features collected from the reading phase as strong indicators of respective parameters.

- Subjective Hatefulness Category: total fixation count, regression count, average fixation count, minimum pupil size, total gaze duration.
- Confidence Ratings: mean fixation count, minimum pupil size, variation in pupil size, total regression-out count.
- Individual Variation: no gaze correlates.

We use these statistical correlations to explain why gaze infusion enhances model performance on HSD (see Section 5.2). We expect gaze data to improve the model's representation of confidence. Since gaze features correlated with subjective hatefulness evaluation and confidence are encoded across all instances, we expect these models being less sensitive to the reductions in the quadrants focusing on the following question:

RQ2 Does incorporating gaze signals change model behaviors in regards to different quadrants?

5.2 Gaze-infused Subjective HSD

As the predictors of subjective hate and confidence, we inject five gaze features[11] on each instance using Wang et al. (2024)'s gaze-infused BERT. Simi-

[10] The gaze features used in this analysis are recorded while the participant reads the statement, before they rate for confidence or hatefulness.

[11] Total fixation count, minimum pupil size, total regression count, mean fixation count, and gaze duration.

lar to text-only counterparts, we conduct our analysis on two model variations: *bert-gaze* and *hsd-gaze*. Table 2 shows the F1 scores of the models on the fixed reduction condition.

In line with expectations, reducing Q-I instances do not have an impact on the gaze infused models (**B3**). Unlike *bert-text*, task-specific fine-tuning (*hsd-text*, *hsd-gaze*) and gaze infusion (*bert-gaze*, *hsd-gaze*) allow models to learn aspects of hate speech, subjectivity, and confidence through the remainder of the instances. The models behave differently on Q-II and Q-III depending on their pretrained model. It should be highlighted that the commonality of these quadrants is uncertainty (low confidence) scores. Both models suffer when Q-III instances (collective uncertainty) are limited (**B2**), though the effect is particularly pronounced in the latter. The subjectivity and confidence infused by these instances appear confusing for this model. HSD pretraining together with gaze infusion results in an overall higher performance, this combination is sensitive to the Q-II (subjective uncertainty) instances, indicating this model benefits from the noisy instances to perform better.

The performance with distance-based reduction drops substantially for the gaze infused models (see Appendix A.7 for the details). This indicates that gaze-infused model learn from the extreme points, and reducing their weights impairs the training.

Backbone/ Pretrained Model	***Equal*** **All**	***Fixed*** **Q-I**	**Q-II**	**Q-III**
bert-gaze	**.627**	.613–	.601↓	.591↓
hsd-gaze	.694	.683 –	.637↓↓	**.701**–

Table 2: Average F1 scores of gaze infused models. Please refer to Table 1 for the details.

Overall, the findings show that different model variants exhibit distinct sensitivities across quadrants (**RQ1**), confirming that these quadrants differ both conceptually and practically. Additionally, models without gaze supervision (Section 4), which do not capture subjectivity, display different patterns from those with gaze supervision (**RQ2**). Fine-tuning and gaze infusion lead models to behave differently across quadrants by enabling them to learn different data patterns. These differences are especially clear through our diagnostic tool, which reveals the models' varying sensitivity to confidence and individual variation. These results highlight two key insights: (i) understanding data characteristics before any data cleaning is essential – what seems like noise may be key to performance (Q-II vs. Q-I, Q-II vs. Q-III) and (ii) enriching the training (e.g. HSD pretraining, gaze infusion) influences the model's learning patterns – instances confusing for one model may be informative for another.

6 Conclusion

In this paper, we showed that human label variation in hate speech detection can come from different sources: uncertainty and genuine subjectivity. We examined how these types of variation can be disentangled using annotator's confidence and demonstrate the importance of having continuous scale ratings during data collection, showing their substantial effect on understanding variation. Our results confirm that the instances in different areas of the subjectivity-uncertainty space (i.e. quadrants) exhibit distinct impacts on model performance, representing conceptually different types of variation (**RQ1**). Our study also highlights the potential of moving from text-only models for hate speech detection, to models that incorporate behavioral features such as gaze. Importantly, these features cannot only serve as significant predictors of subjective opinion (Alacam et al., 2024; Yang and Hollenstein, 2023; Mishra et al., 2016), but serve as key indicators of annotator confidence. While HSD pretraining enhances the performance as expected, gaze-infusion, which implicitly captures both subjective hate evaluation and annotator's confidence, improves the models over their text-only counterparts (particularly for the bert-text model) and changes the model behavior (**RQ2**).

This paper lays the foundational work showing that gaze integration (specifically, significant gaze predictors of hatefulness and confidence scores as components of subjective variation) offers a way of making models aware of different sources of variation, without explicit labeling of these different sources. We therefore believe that gaze is a natural source of data worth exploring to develop models that can "naturally" learn to deal with subjectivity and certainty at the same time.

Limitations

While our approach is designed to be both dataset- and model-agnostic, our current validation focuses on a single hate speech detection dataset. This initial focus provides a strong foundation, and future

extensions to additional opinion-mining tasks and datasets will further demonstrate the generalizability of our method. Although many studies collect confidence ratings, such information is often excluded from final datasets or simplified into binary formats. As the availability of datasets with fine-grained confidence scores continues to grow, we plan to broaden our evaluations of our approach.

The annotator pool in the dataset we used could benefit from increased diversity, strengthening the applicability of our approach across different demographic and socio-cultural groups.

Our train-model-per-annotator approach effectively accounts for subjectivity in hate speech detection but operates with a relatively limited number of training instances per annotator. Yet, we conducted extensive parameter-tuning (including the number of frozen layers), while observing the training and validation loss. Based on these experiments, we are confident in the effectiveness of the models' training despite the limited dataset size.

While it can be argued that fine-grained Likert scoring is more resource-intensive than simpler approaches such as binary labeling, this may limit its suitability for certain research contexts. In the case of subjective evaluations – such as opinions – finer scales offer increased sensitivity to subtle variations in judgment, which is the primary focus of our study. As such, they are preferable for capturing this type of nuanced information.

The original GAZE4HATE dataset is preprocessed with basic signal-level noise reduction during gaze data collection and extraction. The original dataset reports trial-level drift correction and follows the standard SR-EyeLink filters during the data export. In this paper, we use the existing data as is and do not apply any filtering on the gaze data for the sake of reproducibility.

This paper does not aim to provide state-of-the-art HSD performance, but rather to have a deeper understanding of the subjective variation and uncertainty. The BERT-like model architecture is selected due to its already existing gaze-infusion implementation in the field (Wang et al., 2024; Deng et al., 2024; Alacam et al., 2024), enabling a comparative analysis of the impact of gaze (that reflects subjectivity) to text-only models (Section 5). However, any BERT-like architecture could be used in plug-and-play fashion for both binary and multiclass classification. To diagnose text-only models (Section 4), the only criterion is to be able to include sample weighting during the finetuning process, which remains (for the time being) a less straight-forward method for recent large language models. We aim to explore preference optimization in the reward model in future experiments.

Ethics Statement

This study examines subjective variation and uncertainty in hate speech annotation using an existing dataset that contains potentially harmful language along with annotator metadata. Given the sensitive nature of the data and the inherent biases in annotation, our study aims a better understanding of the subjectivity and uncertainty involved, ultimately informing fairer data practices in this area. The dataset includes annotator metadata, such as behavioral features and demographic characteristics (e.g. gaze data, gender and age). However, all annotator identities remain fully anonymized, and no personally identifiable information is included in our analysis. The dataset was made available upon request under ethical guidelines established by its curators. We adhere strictly to these guidelines, comply with all associated terms of use. This paper was proofread with the assistance of an AI grammar checker.

Acknowledgments

The authors (OA, SH, SZ) 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). BP acknowledges funding by ERC Consolidator Grant DIALECT 101043235.

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.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

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.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. Proceedings of Machine Learning Research.

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

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

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

Federico Cala, Pietro Tarchi, Lorenzo Frassineti, Mustafa Can Gursesli, Andrea Guazzini, and Antonio Lanata. 2023. Eye-tracking correlates of the implicit association test. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2023:1–4.

Tommaso Caselli and Flor Miriam Plaza-del Arco. 2025. Learning from disagreement: Entropy-guided few-shot selection for toxic language detection. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 53–66, Vienna, Austria. Association for Computational Linguistics.

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

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

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.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–224, Bangkok, Thailand. Association for Computational Linguistics.

Andy Field, Zoe Field, and Jeremy Miles. 2012. *Discovering statistics using R*. SAGE.

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.

Bruce C Hansen, Pamela J Rakhshan, Arnold K Ho, and Sebastian Pannasch. 2015. Looking at others through implicitly or explicitly prejudiced eyes. *Visual Cognition*, 23(5):612–642.

Sanne Hoeken, Ozge Alacam, Dong Nguyen, and Sina Poesio, Massimo Zarrieß. 2025. Not just who or what: Modeling the interaction of linguistic and annotator variation in hateful word interpretation. In *Proceedings of the 16th International Conference on Computational Semantics (IWCS)*, Düsseldorf, Germany.

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.

Uthman Jinadu and Yi Ding. 2024. Noise correction on subjective datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5385–5395, Bangkok, Thailand. Association for Computational Linguistics.

Ian Krajbich, Carrie Armel, and Antonio Rangel. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13:1292–1298.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles Lima Sanches, Olivier Augereau, and Koichi Kise. 2018. Estimation of reading subjective understanding based on eye gaze analysis. *PloS one*, 13(10):e0206213.

Sebastian Loftus, Adrian Mülthaler, Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2025. Using LLMs

and preference optimization for agreement-aware HateWiC classification. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 538–547, Vienna, Austria. Association for Computational Linguistics.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is LLM an overconfident judge? unveiling the capabilities of LLMs in detecting offensive language with annotation disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5609–5626, Vienna, Austria. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 147–154. IEEE.

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.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

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.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

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.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

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.

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.

Shinsuke Shimojo, Claudiu Simion, Eiko Shimojo, and Christian Scheier. 2003. Gaze bias both reflects and influences preference. *Nature neuroscience*, 6(12):1317–1322.

Giora Simchoni and Saharon Rosset. 2023. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Armin W Thomas, Felix Molter, Ian Krajbich, Hauke R Heekeren, and Peter NC Mohr. 2019. Gaze bias differences capture individual choice behaviour. *Nature human behaviour*, 3(6):625–635.

Alexandra 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.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Bingbing Wang, Bin Liang, Lanjun Zhou, and Ruifeng Xu. 2024. Gaze-infused bert: Do human gaze signals help pre-trained language models? *Neural Computing and Applications*, pages 1–22.

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 Konvens*.

Duo Yang and Nora Hollenstein. 2023. Plm-as: Pre-trained language models augmented with scanpaths for sentiment classification. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 4.

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.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Appendix

A.1 Supplementary Information on Data

The hateful sentences in the GAZE4HATE dataset originate from the FEMHATE dataset[12], which compiles annotations from multiple annotators of different genders. These sentences exhibit variation in the degree of hatefulness, as discussed in (Wojatzki et al., 2018).

Unlike the analyses conducted by (Alacam et al., 2024) that focus solely on the reading phase, our disentanglement approach incorporates gaze features from all four experimental phases: sentence reading, hatefulness rating, confidence rating, and rationale annotation. Aligning data at the instance level requires complete gaze recordings from all phases. However, missing gaze values, particularly during the confidence and hatefulness rating phases, pose challenges for statistical modeling. To address this, we exclude instances with missing values in any phase (N = 267). The final dataset used in this study consists of 3,360 instances with aligned gaze and behavioral data from all four phases[13].

Additionally, *a priori* labels in this dataset comprise six distinct categories: strongly hateful towards women, mean statements towards women, strongly hateful towards men, empowering statements about women, neutral statements about women, neutral statements unrelated to gender. These categories are later consolidated into three *a priori* labels as hateful, neutral and positive. It should be highlighted that in this study instead of predicting a priori label, we focus on subjective hate ratings provided by users.

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), each annotating 90 sentences. They were paid or given a course credit to participate. The experiment took approximately 40 minutes for each participant. for the details of experimental setup, please refer to the original dataset paper (Alacam et al., 2024) .

A.2 Supplementary Explanatory Analysis on the components of subjective variation

We utilize sentence and annotator entropy measures to check the existence of extreme outliers based on setting a threshold on z-score transformed entropy. We set the z-score threshold ± 3.29 since instances beyond these values are commonly treated as extreme outliers (Field et al., 2012).

Figure 5 shows the heatmap (bottom part) for hatefulness score on each instance. X-axis corresponds to sentences, whereas y-axis denotes the annotators. Top part visualizes the sentence entropy values.

In our study, sentence entropy quantifies the purity of the sentence in terms of evaluated hatefulness categories. The more the annotators rate the item in the same category, the lower the sentence entropy (darker colors in Figure 5, that visualizes the high variation in the hatefulness ratings across participants and sentences).

For the right side of the cartesian plane shown in Figure 3, we observe instances with higher sentence entropy, whereas they are not frequent on the left side of the plane (associated with lower confidence). This indicates that annotators who distribute labels evenly across hate, neutral and positive classes tend to exhibit higher confidence

[12] https://github.com/muchafel/femhate

[13] During aligning the instances among different experimental phases (sentence reading, hatefulness rating, confidence rating, and rationale annotation), the instance is removed if it contains any missing values in any experimental phase (N = 267) as detailed in Appendix A.1)

in their assessments.

Although sentence entropy – as commonly used metric – can be used to obtain candidate sentences with variation, this approach alone does not give further information whether the variation is observed due to subjectivity or uncertainty. The sentence entropy (z-score transformed) of all sentences lie between ± 2 z-score (from -1.84 to 1.57) indicating no particular outlier at the sentence level.

We also calculate annotator entropy (that quantifies the class distribution of the labels given by the same annotator), where lower entropy indicates a stronger tendency toward a particular class.

As illustrated in detail in Appendix Figure 5 some annotators exhibit a preference for assigning higher hatefulness ratings (e.g., P6, P43), while others tend to give more neutral ratings (e.g., P35) or more positive ratings (e.g., P32).

The annotator attributes (gender). The annotator attributes (similar to sentence attributes) could serve to understand the dataset and model's tendencies when fitted to the proposed concept. GAZE4HATE paper (Alacam et al., 2024) reports that there is no statistically significant difference in the hatefulness ratings between female and male gender categories. For the confidence scores, male participants indicate higher confidence in evaluating positive or neutral statements compared to female participants, although this difference is not significant. In terms of confidence for hateful statements, there is no difference between these two gender categories at all. When we added the gender into the base GLMs introduced in Section 5, it did not provide any statistically significant contribution in any of these three settings, confirming the original paper's results.

Here we additionally provide the distribution of evaluations on the quadrants based on the participant's gender (Table 3). Q-IV (collective opinion) is the most populated quadrant by both genders, and Q-II (subjective uncertainty) is the least populated quadrant. Male participants slightly favored Q-III (collective uncertainty) over Q-I (subjective opinion); a more pronounced reversed effect was observed for the female participants.

Variation in confidence dimension. Due to space limitations, in this paper, we focused primarily on the variation in the hate-speech dimension, while leaving the detailed analysis of annotator confidence variation aside. However, we conducted a similar explanatory analysis on the confidence variation (as in Figure 5), it is safe to say that there are some annotators who exhibit less confidence overall (41, 6) or high confidence (25,36). We applied z-score transformation to all of our measures at the participant level.

	Q-I	Q-II	Q-III	Q-IV
female	27.2%	15.8%	22.1%	34.9%
male	24.4%	19.0%	26.5%	30.1

Table 3: The distribution of evaluations on the quadrants based on the participant's gender.

A.3 Features

META information

- annotator ID
- Sentence ID
- *A priori* label multi: hate, neutral, positive
- *A priori* label binary: hate, nohate
- Gender of the annotator
- Age of the annotator
- Ling type: whether the sentence connotation is made explicitly or implicitly

Gaze Features information (MEAN, MAX, MIN, SUM)

- Min Pupil: Min. Fixation Pupil Size on AOI
- FC: Fixation Count on AOI
- RC: Run count on AOI
- RIC: Regression In Count on AOI
- ROC: Regression Out Count
- DT: Dwell Time on AOI
- FFD: First Fixation Duration
- Var Pupil: Pupilsize variation
- TDT: TRIAL DWELL TIME
- TFC: TRIAL FIXATION COUNT

Annotation Features

- Intensity rating by each annotator in 1-to-7 Likert Scale (*Intensity rating*)
- Multiclass Hatefulness category (Intensity Category)
- Binary Hatefulness category based on intensity ratings (Intensity Category Binary)
- Confidence rating by each annotator in 1-to-5 Likert Scale (Confidence rating)
- whether the token is clicked or not (Clicked)

A.4 Statistical Models

Subjective Hatefulness Category The significant predictors of the subjective hate ordered by the

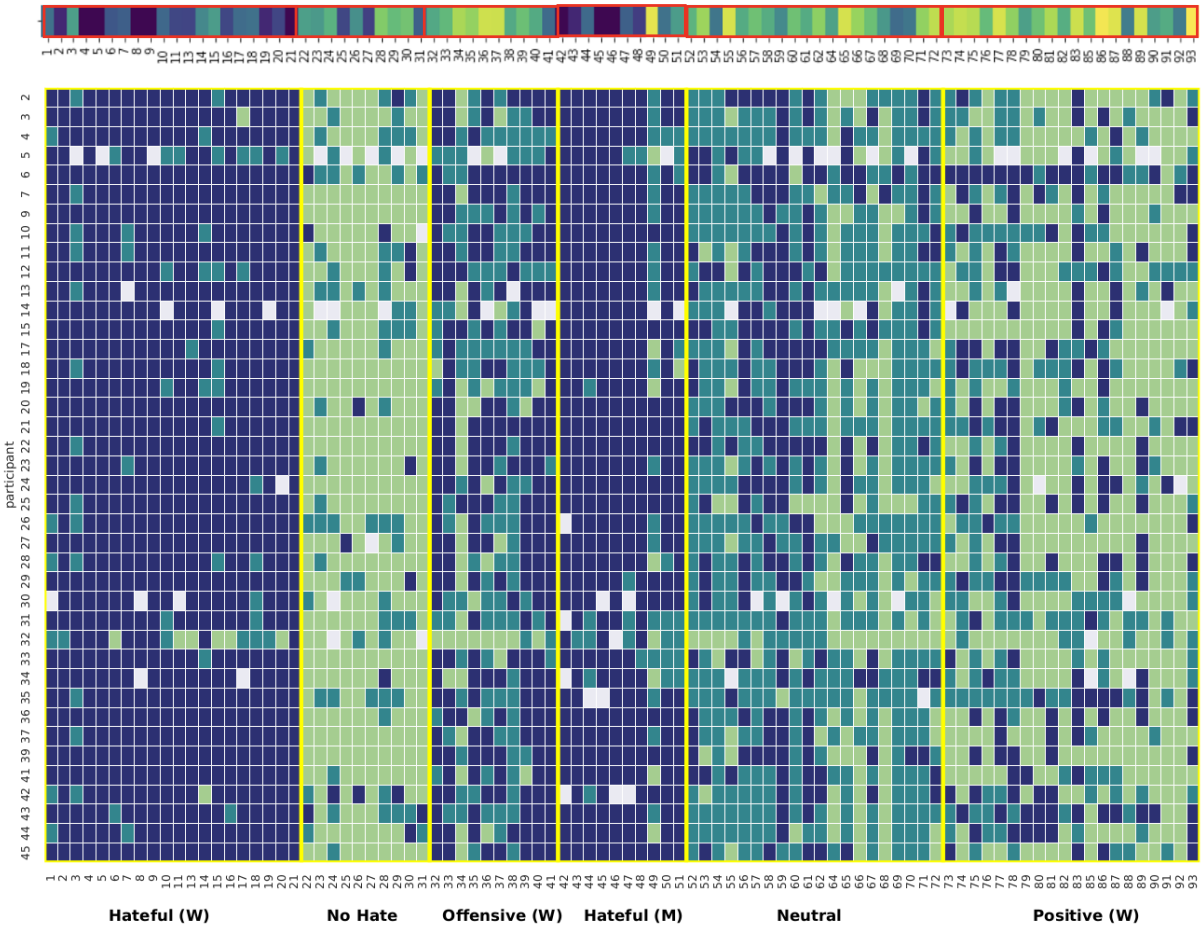

Figure 5: Top: Sentence Entropy (x-axis represents the sentences). Yellow color denotes high entropy, low purity in the class distribution. Bottom: Annotator-Sentence Hatefulness Category Heatmap (navy: hateful, dark green: neutral, light green: positive).

magnitude of their estimates: total fixation count, regression count, average fixation count, minimum pupil size, total gaze duration, average pupil size, regression-in count, maximum saccade amplitude (in reading), pupil size variation, maximum saccade velocity (in reading), average gaze duration.

$$\begin{aligned} base =& glmer(hate\_cat. \sim (1|part.)) \\ model_i =& glmer(hate\_cat. \sim ind\_var_i \\ & + (1|part.)) \end{aligned}$$

First, we add the features individually to the base model that only has random effects to measure their individual effects on the hatefulness evaluation. The model comparison of the model with the particular variable to the base model is performed using ANOVA.

Confidence Ratings. We fit a Cumulative Link Models for Ordinal Regression with 32 independent features (one at a time) and 2 random effects (as annotator and sentence). Figure 7 (top) presents the distribution of the z-score transformed confidence ratings.

$$\begin{aligned} base =& clmm(conf\_rate \sim (1|part.) + (1|sent.)) \\ model_i =& clmm(conf\_rate \sim ind\_var_i \\ & + (1|part.) + (1|sent.)) \end{aligned}$$

Individual Variation. We fit a mixed effect linear model with 32 independent features (one at a time) and two random effects (as annotator and sentence). Figure 7 (bottom) presents the distribution of the z-score transformed individual variation.

Nine predictors showed significant effect on the individual variation: sentence entropy, maximum saccade velocity (in conf. rating), confidence rating, maximum saccade amplitude (in conf. rating), maximum saccade amplitude (in hate rating), peak saccade velocity (in conf. rating), average saccade amplitude (in hate rating), clicked token ratio and intensity rating.

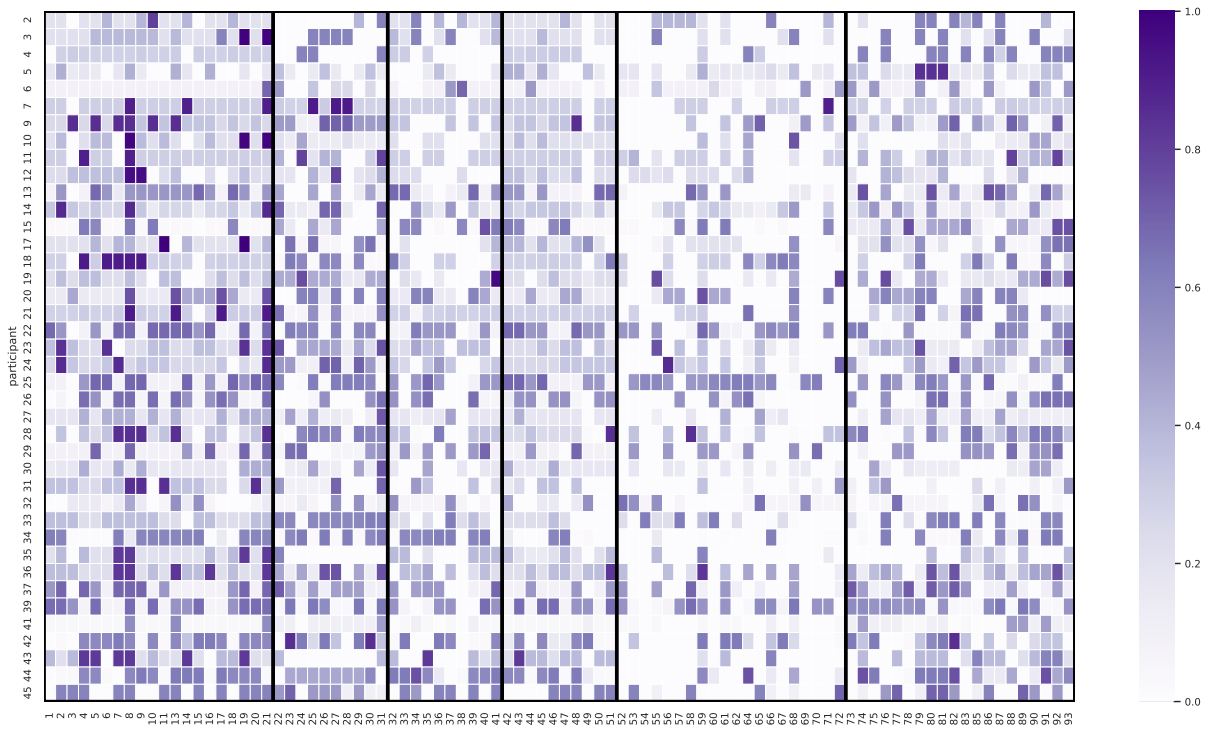

Figure 6: Annotator-confidence score heatmap. X-axes denotes the sentences. Darker purple corresponds to higher confidence.

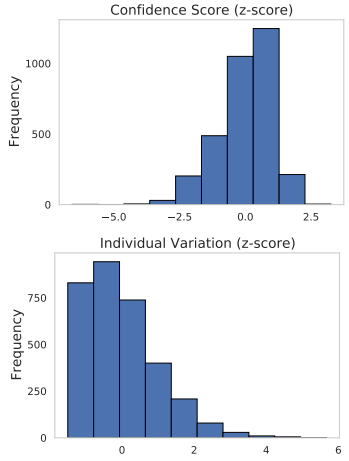

Figure 7: Distribution of Confidence Score (top) and Individual Variation (bottom) parameters (z-score–transformed)

$$base = lmer(individual\_variation. \sim (1|sent.))$$
$$model_i = lmer(individual\_variation \sim indep.\_feat_i + (1|sent.))$$

A.5 Class Distributions per Quadrant

In order to facilitate qualitative analysis on the disentanglement method, the frequencies of observations for each parameters per quadrant are summarized in Table 9.

Train-val-test splits. The provided train-val-test splits in the GAZE4HATE dataset are created in a way that there is no overlap among the splits in terms of sentence. On the other hand, our approach requires a finer-grained splitting strategy focusing on the quadrants. The projected instances of same sentence could be distributed to all quadrants, i.e. same sentence can be evaluated with less certainty by one annotator, but with certainty by another (or with varying variation by another. Thus, we decided to utilize individualized models as explained in the main paper.

The quadrant-driven splitting method also ensures that the final splits for each annotator have enough (and proportional) data from each quadrants. Since we keep the proportion of each quadrant per annotator while creating train-val-test splits, the individualized models still aligns with the annotator' tendency, without introducing additional class imbalance noise as noted in (Mostafazadeh Davani et al., 2022).

A.6 Supplementary: Models

Given the individualized nature of the models, each is trained on a relatively small amount of obser-

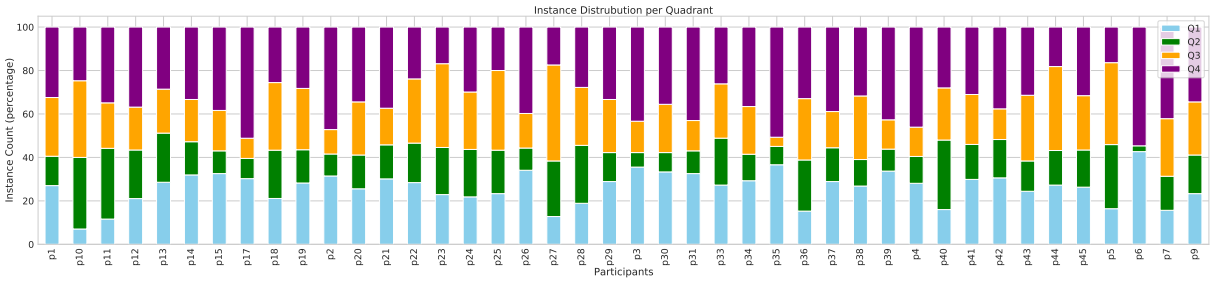

Figure 8: Number of instances in each quadrant per annotator.

QII (N=570)	QI (N= 881)
Sent entropy: High: 328, 171: 2, Low: 71	Sent entropy: High: 350, Med: 302, Low: 229
Part entropy: High: 468, Med: 100, Low: 2	Part entropy: High: 680, Med: 169, Low: 32
Ling type: Explicit: 307, Implicit: 263	Ling type: Explicit: 476, Implicit: 405
RT categories: Fast: 277, Slow: 293	RT categories: Fast: 663, Slow: 218
Intensity: Hate: 267, Positive: 142, Neutral: 161	Intensity: Hate: 410, Positive: 340, Neutral: 131
A priori label: Hate: 225, Positive: 231, Neutral: 114	A priori label: Hate: 342, Positive: 403, Neutral': 136
Distance (z-score): Max: 1.0 , Mean: 0.22, SD: 0.13	Distance (z-score): Max: 0.70 , Mean: 0.18 , SD:0.11
QIII (N=775)	**QIV (N=1137)**
Sent entropy: High: 407, Med: 267, Low: 101	Sent entropy: High: 305, Med: 423, Low: 409
Part entropy: High: 619, Med: 156, Low: 0	Part entropy: High: 885, Med: 211, Low: 41
Ling type: Explicit: 426, Implicit: 349	Ling type: Explicit: 627, Implicit: 510
RT categories: Fast: 416, Slow: 359	RT categories: Fast: 867, Slow: 510
Intensity: Hate: 313, Positive: 175, Neutral: 287	Intensity: Hate: 626, Positive: 219, Neutral: 292
A priori label: Hate: 306, Positive: 244, Neutral: 225	A priori label: Hate: 608, Positive: 252, Neutral: 277
Distance (z-score): Max: 0.56 , Mean: 0.18, SD: 0.09	Distance (z-score): Max: 0.35 , Mean: 0.14, SD: 0.06

Figure 9: The class distributions for seven variable for each Confidence-Individual Variation Quadrant

vations (with a maximum of 90 sentences). To optimize the performance, we have experimented with freezing the first 6, 8, 10 layers of the BERT model and compared to a fully trainable model. Evaluation metrics indicate that freezing the first six layers yields best performance, thus we adopt this configuration for the rest of the analysis.

A.7 Distance-based Reduction for Gaze-infused HSD

As summarized in Table 4, while extreme cases may introduce noise for the text-only models, the variation captured in gaze features is crucial for learning different notions of variation for the gaze-infused models.

For the gaze-infused hsd model, excluding extreme instances from Q-I has substantial negative effect on the performance. There are notable differences in the impacts of different quadrants on subjective hate speech task. This table effectively illustrates that when additional information reflecting subjectivity and uncertainty is incorporated, the models become more sensitive to extreme points (bert-gaze). However, if the model has already undergone extensive pretraining on hate speech detection (hsd-gaze), reducing the weight of these instances diminishes this added value, leading to lower performance scores.

Backbone Pretrained Model	*Equal* All	*Distance-based* Q-I	Q-II	Q-III
bert-gaze	**.627**	.597↓	.556↓↓	.618–
hsd-gaze	**.694**	.636↓↓	.680–	.648 ↓↓

Table 4: F1 scores with and without HSD fine-tuning, where the lowest value per model is underlined and the highest values marked as bold. $\Delta <= .02p$ indicates negligible or no change (–). ↓ or ↑ signifies $\Delta > 0.02p$. ↓↓ or ↑↑ indicates a higher difference, $\Delta > 0.05p$. Here Δ is the absolute difference between the respective score and performance of the same model trained with equal weights.

A.8 Training Details and Parameter Optimization

We conducted extensive parameter-tuning (including the number of frozen layers), while observing the training and validation loss. Based on these experiments, we are confident in the effectiveness of the individual model's training despite the limited dataset size.

Training was performed on a NVIDIA® RTX™ A6000 (48 GB). Parameters are optimized with GridSearch on the following sets.

- learning rate : 1e-5, 3e-5, 5e-5, 1e-4, 3e-4
- batch sizes : 10, 20, 50, 100, 200
- number of epochs : 20, 50
- number of frozen layers (first) : 0, 6, 8, 10

Based on the results, the individualized models are trained using lr: 1e-5, batch size: 10, epochs:20 and freezing the first 6 Bert layers.

A.9 Weight Reduction

The training size was kept constant across all settings (no removal of instances). In order to check whether the change in the performance is originated from the characteristics of the downgraded data instances rather than the number of the instances being manipulated, we conducted the following analysis.

	Equal All	Q-I	Q-II	Q-III	Q-IV	17%	26%	34%
bert-text	.574	.544	.556	.592	.570	.611	.644	.571
hsd-text	.679	.675	.671	.649	.632	.662	.637	.697

Table 5: Extended Results for Table 1 on the fixed-reduction setting with random instance selection.

When we systematically lower the weights of all instances in the dataset to varying degrees, we do not only observe drop in the performance but also the improvements. As an example, *bert-text* model (using distance-based measures) in Table 1 exhibits increases in the performance since extreme cases have less effect.

Moreover, we conducted two additional manipulations. First, we reduced the weights of the instances in the Q-IV (as the most populated quandrant) that denotes to "collective opinion". Second, we randomly reduced P% of instances from each quadrant (where P = 17, 26, 34, P corresponds to the ratio of data manipulated, selected based on the ratio of instances in the quadrants to the full data size: e.g. Q-II contains 17% of all instances, Q-IV contains 34%, and 26% for Q-II and Q-III on average separately (see Table-8). As shown in Table 5, our findings indicate that reducing the weight of Q-IV instances does not impact *bert-text* performance. However, for *hsd-text*, it leads to a decrease in performance, similar to Q-III, also with low variation. On the other hand, random reductions from all quadrants in different proportions lead to both performance increases and decreases. Overall, these results support our conclusion that the observed effects stem from the characteristics of the instances rather than the number of instances with downgraded weights.