Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks

Artem Vazhentsev[1,2], Gleb Kuzmin[1,4], Akim Tsvigun[5,8], Alexander Panchenko[2,1], Maxim Panov[6], Mikhail Burtsev[7], and Artem Shelmanov[3]
[1]AIRI, [2]Skoltech, [3]MBZUAI, [4]FRC CSC RAS, [5]AI Center NUST MISiS, [6]TII,
[7]London Institute for Mathematical Sciences, [8]Semrush
{vazhentsev, kuzmin, panchenko}@airi.net
maxim.panov@tii.ae artem.shelmanov@mbzuai.ac.ae

Abstract

Many text classification tasks are inherently ambiguous, which results in automatic systems having a high risk of making mistakes, in spite of using advanced machine learning models. For example, toxicity detection in user-generated content is a subjective task, and notions of toxicity can be annotated according to a variety of definitions that can be in conflict with one another. Instead of relying solely on automatic solutions, moderation of the most difficult and ambiguous cases can be delegated to human workers. Potential mistakes in automated classification can be identified by using uncertainty estimation (UE) techniques. Although UE is a rapidly growing field within natural language processing, we find that state-of-the-art UE methods estimate only epistemic uncertainty and show poor performance, or under-perform trivial methods for ambiguous tasks such as toxicity detection. We argue that in order to create robust uncertainty estimation methods for ambiguous tasks it is necessary to account also for aleatoric uncertainty. In this paper, we propose a new uncertainty estimation method that combines epistemic and aleatoric UE methods. We show that by using our hybrid method, we can outperform state-of-the-art UE methods for toxicity detection and other ambiguous text classification tasks[1].

1 Introduction

Many natural language processing (NLP) tasks are subjective and contain inherent ambiguity. For example, the notion of toxicity is inherently subjective (Waseem, 2016) and can be defined in a number of ways that may conflict with one another and differ according to the demographic that the methods are applied to (Thylstrup and Waseem, 2020). For many datasets, implicit or ambiguous toxicity can comprise more than 90% of the labeled toxic content (Hartvigsen et al., 2022). Such ambiguity introduces a high risk of classification mistakes for machine learning (ML) models. Classification mistakes for toxicity detection can result in the removal of legitimate non-toxic content on one hand, and the lack of sanction for toxic content, on the other. A common method for addressing this concern for content moderation is to abstain from predictions on ambiguous instances and process them with the help of human workers (Roberts, 2019).

A classification task where some model predictions can be "rejected" is called *selective classification* (Geifman and El-Yaniv, 2017). The common approach to solving it is applying uncertainty estimation (UE) techniques. UE is a field of ML that seeks to model the degree to which model predictions can be trusted by correlating model mistakes and performance. Better UE methods improve the performance of selective classification and the trade-off between the amount of labor and the reliability of downstream applications. In toxicity detection, better UE methods minimize the amount of content that is reviewed by human moderators to predominately be classification errors.

Recent works have suggested deterministic approaches to UE of neural network predictions based on fitting the density of latent instance representations (Lee et al., 2018; van Amersfoort et al., 2020; Mukhoti et al., 2023; Yoo et al., 2022; Kotelevskii et al., 2022). They have shown good performance in NLP for the detection of out-of-distribution (OOD) instances, adversarial attacks, and misclassified objects in non-ambiguous tasks. However, they primarily capture epistemic uncertainty, i.e. uncertainty related to the lack of knowledge about model parameters and training data, overlooking aleatoric uncertainty, i.e. uncertainty that arises from ambiguity and noise in data.

This work aims to create a UE method for more reliable selective classification in ambiguous tasks such as toxicity detection by combining different

[1]The code for reproducing experiments is available online at https://github.com/AIRI-Institute/hybrid_uncertainty_estimation

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11659–11681
July 9-14, 2023

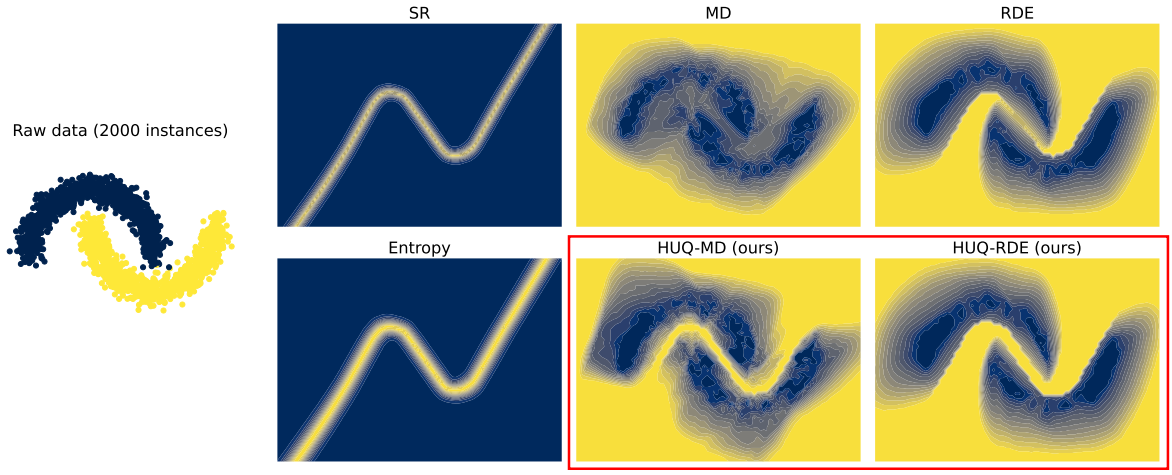

Figure 1: The left image shows training data from the two moons dataset. The second column shows uncertainty scores obtained using aleatoric UE methods entropy and Softmax Response (SR). The first row of the two rightmost columns illustrates scores obtained using epistemic UE methods Mahalanobis Distance (MD) and Robust Density Estimation (RDE) (see Section 3.2), and the second row shows the scores obtained using our method (HUQ). The lighter color indicates higher uncertainty. HUQ correctly identifies both regions with untrustworthy predictions: the area away from the training data distribution and the area around the model decision boundary.

types of uncertainty. Instances that carry a high risk of classification mistakes come from two sources: a) OOD areas, which can be detected with epistemic UE methods; and b) in-distribution ambiguous areas, for detection of which, aleatoric UE methods are appropriate (for illustration see Figure 1). Therefore, we propose a Hybrid Uncertainty Quantification (HUQ) method that switches between epistemic and aleatoric uncertainties or linearly combines them. It produces better scores of total uncertainty, which subsequently leads to better selective classification. The experiments on various ambiguous tasks show that HUQ in a majority of cases significantly outperforms other state-of-the-art UE techniques. To summarize, the contributions of this work are the following.

- In Section 4, we propose a new uncertainty estimation method *HUQ* that combines epistemic and aleatoric UE techniques in a special way that allows to improve the quality of selective classification in ambiguous tasks.
- To the best of our knowledge, this work is the first to conduct an empirical investigation of state-of-the-art UE methods for ambiguous text classification tasks such as toxicity detection. Our analysis shows that the proposed HUQ approach outperforms state-of-the-art methods in selective text classification on ambiguous tasks; see Sections 5 and 6.
- We analyze the limitations of the proposed method and suggest conditions to be met for achieving the improvements; see Section 7.

2 Related Work

Quantifying uncertainty of deep neural network predictions can be successfully accomplished using deep ensembles (DE; Lakshminarayanan et al., 2017), Bayesian models (Blundell et al., 2015), or their approximations. However, most of these methods have various drawbacks, including large computational overhead. For example, for DE, we need to multiply training time, the occupied memory, and inference time, since this network requires training, storing, and running inference for multiple versions of the same model. This makes DE hardly applicable in real-world scenarios.

Recent work has investigated computationally efficient deterministic approaches (e.g., Lee et al., 2018; van Amersfoort et al., 2020; Liu et al., 2020). However, most work is based on feature space density and focuses only on the OOD detection task and epistemic uncertainty estimation. Another computationally efficient approach is SelectiveNet (Geifman and El-Yaniv, 2019), which was designed for computer vision tasks. It introduces two separate heads for prediction and selection within the model architecture and adds a special loss component to minimize selective risk with a specified coverage.

Most similar to our work is Mukhoti et al. (2023), which also considers both aleatoric and epistemic uncertainty. DDU uses a combination of feature-

space density for epistemic uncertainty and the softmax predictive distribution for aleatoric uncertainty. They advocate for the usage of different methods for quantifying uncertainty, depending on whether a considered instance is ID or OOD. However, they overlook using a linear combination of uncertainty scores, relying solely on feature-space density for instances considered OOD. We note that these instances can also be borderline (instances from middle to low-density areas), for which using aleatoric uncertainty measures may also be appropriate. Besides, Mukhoti et al. (2023) do not provide results for selective classification and mostly experiment with image classification tasks.

Recently, selective classification (or misclassification detection) has been studied for NLP tasks.One line of such work has proposed adding a regularization term to the training loss. Xin et al. (2021) introduces a penalty term for confident instances with a high loss value. Another approach proposed by Zhang et al. (2019) uses a metric regularization that minimizes the inter-class distance in the latent feature space while maximizing the margin between classes. He et al. (2020) propose a regularization technique based on self-ensembling that aims to minimize the difference between predictions of the two versions of the model. They also combine this approach with mix-up (Thulasidasan et al., 2019) and a distinctiveness score based on the MD. Some work has also considered approximations of deep ensembles based on Monte-Carlo dropout (e.g., Shelmanov et al., 2021; Vazhentsev et al., 2022). Vazhentsev et al. (2022) conduct a vast empirical investigation and suggest several promising combinations of regularizers and feature-density-based methods. They also highlight the importance of spectral normalization for obtaining good results. Kotelevskii et al. (2022) propose a new UE method NUQ and test it for text classification models trained in the low-resource regime.

Despite the aforementioned efforts, highly ambiguous text classification tasks such as toxicity detection have been overlooked in the previous work. Moreover, to the best of our knowledge, no prior work in NLP takes into account aleatoric uncertainty and combines multiple types of uncertainty for a holistic view of uncertainty.

3 Background

Two types of uncertainty have been documented in the literature: aleatoric and epistemic (Der Kiureghian and Ditlevsen, 2009). *Aleatoric*, or data uncertainty, arises from ambiguity and noise in data. It should be high, for example, for groups of instances prone to annotation discrepancy. *Epistemic*, or model uncertainty, pertains to a lack of knowledge about model parameters and can often be mitigated through additional training data collection. Epistemic uncertainty is particularly important for OOD detection (Hendrycks and Gimpel, 2017) and active learning (Settles, 2009).

According to the Bayesian approach to measuring uncertainty in deep learning networks (Blundell et al., 2015; Gal, 2016; Depeweg et al., 2018), the total uncertainty of a model prediction $\mathbf{x}$ is a sum of aleatoric $U_{\text{A}}(\mathbf{x})$ and epistemic uncertainty $U_{\text{E}}(\mathbf{x})$:

$$U_{\text{T}}(\mathbf{x}) = U_{\text{A}}(\mathbf{x}) + U_{\text{E}}(\mathbf{x}). \quad (1)$$

High total uncertainty should correlate with classification mistakes and can be used to flag model predictions for human review.

3.1 Out-of-Distribution and (Ambiguous) In-Distribution Instances

We define *out-of-distribution (OOD)* instances $\mathcal{X}_{\text{OOD}}$ as those located either outside a training data distribution or in its low-density regions. They can be identified by high epistemic uncertainty.

In-distribution (ID) instances we define to belong to the domain of the dataset $\mathcal{D}$ located "inside" the training data distribution. ID instances are those, for which model predictions have very small epistemic uncertainty, i.e. below some threshold $\delta_{\min}$:

$$\mathcal{X}_{\text{ID}} = \{\mathbf{x} \colon U_{\text{E}}(\mathbf{x}) \leq \delta_{\min}\}. \quad (2)$$

Note that for the in-distribution data, on the basis of (1) and taking into account (2), we can empirically approximate $U_T(\mathbf{x}) \simeq U_{\text{A}}(\mathbf{x})$.

We also define *ambiguous in-distribution (AID)* instances as those, predictions on which having the highest values of aleatoric uncertainty with a lower bound $\delta_{\max}$. AID instances lie around the class-decision boundaries virtually established by the discriminative model:

$$\mathcal{X}_{\text{AID}} = \{\mathbf{x} \in \mathcal{X}_{\text{ID}} \colon U_{\text{A}}(\mathbf{x}) > \delta_{\max}\}. \quad (3)$$

3.2 Quantifying Epistemic Uncertainty

Recent works have proposed a variety of computationally efficient methods for quantifying epistemic uncertainty on the basis of fitting the probability density of latent instance representations. In

this work, we experiment with Mahalanobis Distance (MD, Lee et al., 2018), Robust Density Estimation (RDE, Yoo et al., 2022), and Deep Deterministic Uncertainty (DDU, Mukhoti et al., 2023).

Let $\mathcal{D}$ be a training dataset, $h(\mathbf{x})$ be a latent representation of an instance $\mathbf{x}$ (it is usually taken from the penultimate layer of the network), and $c \in C$ be a class. The UE method based on MD (Lee et al., 2018), for each class, fits a Gaussian centered in a class centroid $\{\mu_c\}_{c \in C}$ with a covariance matrix Σ shared across classes. The highest class-conditional probability density $p(h(\mathbf{x}) \mid y = c)$ determines the confidence of the prediction, and the uncertainty score is computed as the Mahalanobis distance between $h(\mathbf{x})$ and the closest centroid:

$$U_{\text{E}}^{\text{MD}}(\mathbf{x}) = \min_{c \in C} (h(\mathbf{x}) - \mu_c)^T \Sigma^{-1} (h(\mathbf{x}) - \mu_c).$$

RDE (Yoo et al., 2022) improves on MD by computing the covariance matrix Σ_c for each individual class using the Minimum Covariance Determinant estimation (Rousseeuw, 1984) and by reducing the dimensionality of the hidden representations via PCA decomposition with an RBF kernel. These modifications aim to minimize the determinant of the covariance matrix and reduce the influence of outliers in the training data.

DDU (Mukhoti et al., 2023) fits a Gaussian Mixture Model (GMM) $p(h(\mathbf{x}), y)$ with a single mixture component per class. The uncertainty score is the probability density of $h(\mathbf{x})$ under the GMM:

$$U_{\text{E}}^{\text{DDU}}(\mathbf{x}) = \sum_{c \in C} p(h(\mathbf{x}) \mid y = c)\ p(y = c),$$

where $p(h(\mathbf{x}) \mid y = c) \sim \mathcal{N}(h(\mathbf{x}) \mid \mu_c, \Sigma_c)$ and $p(y = c) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \mathbf{1}[y_i = c]$.

Methods based on the fitting density of latent representations are suitable for finding OOD instances but are not capable of identifying AID instances. More generally, they are not good estimators of uncertainty in $\mathcal{X}_{\text{ID}}$. Therefore, for ambiguous tasks where AID instances comprise a large portion of the data, these epistemic UE methods cannot fully cover all potential misclassifications.

3.3 Quantifying Aleatoric Uncertainty

As measures of aleatoric uncertainty, we use two well-known methods based on probabilities from the output softmax layer of a neural network: entropy (Gal, 2016) and Softmax Response (SR, Geifman and El-Yaniv, 2017):

$$U_{\text{A}}^{\text{Ent}}(\mathbf{x}) = -\sum_{c \in C} p(y = c \mid \mathbf{x}) \log p(y = c \mid \mathbf{x}),$$

$$U_{\text{A}}^{\text{SR}}(\mathbf{x}) = 1 - \max_{c \in C} p(y = c \mid \mathbf{x}).$$

Entropy and SR have been proposed also as measures of total uncertainty (Malinin and Gales, 2018). However, this assumption holds only when one has access to the full posterior distribution under the Bayesian paradigm, i.e. all possible uncertainties are quantified within the model. In practice, training datasets are limited, and we can only approximate considered probability distributions. Thus, these methods could not capture all the epistemic uncertainty and mostly reflect the aleatoric one (van Amersfoort et al., 2020; Mukhoti et al., 2023).

4 Hybrid Uncertainty Quantification

There are two major sources of mistakes in model predictions: OOD instances and instances that lie in proximity to the decision boundary (AID instances). Aleatoric uncertainty can help to detect AID instances, while epistemic uncertainty can help to detect OOD instances. In many tasks, we have to deal with both types of mistakes arising from task ambiguity or from a marked covariate shift between training and test data. To address this issue, we propose a hybrid method that combines the strengths of aleatoric and epistemic uncertainty.

Our hybrid uncertainty quantification (HUQ) method first uses Eq. (2) to determine whether an instance $\mathbf{x}$ is ID or OOD. If $\mathbf{x} \in \mathcal{X}_{\text{ID}}$, HUQ applies Eq. (3) to determine if $\mathbf{x}$ is near a class-decision boundary, i.e., $\mathbf{x} \in \mathcal{X}_{\text{AID}}$. Once the type of instance has been identified, we can apply an appropriate uncertainty estimation method for it or combine multiple uncertainty scores into a single estimate.

Uncertainty scores from different methods may however not be comparable with one another due to different magnitudes. Therefore, instead of using absolute values, we propose to rank instances in some dataset $\mathfrak{D}$ by their uncertainty scores and as a final score use these ranks or their combinations. Ranking can be considered as a form of normalization. Moreover, such an approach is desirable for the selective classification task, as we are only interested in the ability to rank predictions by their uncertainty. We define a ranking function $R(\mathbf{u}, \mathfrak{D})$ as the rank of $\mathbf{u}$ over a sorted dataset $\mathfrak{D}$, so $\mathbf{u}_1 > \mathbf{u}_2$ implies $R(\mathbf{u}_1, \mathfrak{D}) > R(\mathbf{u}_2, \mathfrak{D})$.

Having the ranks according to epistemic and aleatoric scores and the type of $\mathbf{x}$, we can define the final total uncertainty score. We consider predictions for ID instances as the most trustworthy, therefore, we define their total uncertainty score as the rank of their aleatoric score $R(U_{\text{A}}(\mathbf{x}), \mathcal{D}_{\text{ID}})$ only among known ID instances $\mathcal{D}_{\text{ID}} = \{\mathbf{x}_i \colon \mathbf{x}_i \in \mathcal{D} \cap \mathbf{x}_i \in \mathcal{X}_{\text{ID}}\}$. Predictions on AID instances are considered the most error-prone. Their total score is the rank of the aleatoric score among all known instances $R(U_{\text{A}}(\mathbf{x}), \mathcal{D})$. Lastly, for $\mathbf{x} \notin \mathcal{X}_{\text{ID}}$, we calculate a linear combination of ranks of aleatoric and epistemic scores among all known instances: $(1-\alpha)R(U_{\text{E}}(\mathbf{x}), \mathcal{D}) + \alpha R(U_{\text{A}}(\mathbf{x}), \mathcal{D})$, where $\alpha \in [0, 1]$ is a task-specific hyperparameter that depends on the quality of the softmax classifier, and a number of training instances. The usage of a mixture rather than only the epistemic score is justified by the fact that the generalization capabilities of models allow them to make meaningful predictions also in OOD regions, so aleatoric scores to some extent remain meaningful in these areas.

Thus, the total uncertainty score for $\mathbf{x}$ according to HUQ is

$$U_{\text{HUQ}}(\mathbf{x}) = \begin{cases} R(U_{\text{A}}(\mathbf{x}), \mathcal{D}_{\text{ID}}), \forall \mathbf{x} \in \mathcal{X}_{\text{ID}} \setminus \mathcal{X}_{\text{AID}}, \\ R(U_{\text{A}}(\mathbf{x}), \mathcal{D}), \forall \mathbf{x} \in \mathcal{X}_{\text{AID}}, \\ (1-\alpha)R(U_{\text{E}}(\mathbf{x}), \mathcal{D}) + \\ \alpha R(U_{\text{A}}(\mathbf{x}), \mathcal{D}), \forall \mathbf{x} \notin \mathcal{X}_{\text{ID}}. \end{cases}$$

Note that HUQ can plug-in various "base" methods for the estimation of epistemic and aleatoric uncertainty. Algorithm 1 summarizes the uncertainty score calculation procedure according to HUQ.

The threshold hyperparameters ($\delta_{\min}$, $\delta_{\max}$) that determine $\mathbf{x} \in \{\mathcal{X}_{\text{ID}} \mid \mathcal{X}_{\text{AID}} \mid \mathcal{X}_{\text{OOD}}\}$ can be set using the validation dataset. We set $\delta_{\min}$ to be the epistemic uncertainty score of the instances $\mathbf{x}$ with the lowest $\beta\%$ epistemic uncertainty on the training set. Similarly, the hyperparameter $\delta_{\max}$ is selected as the uncertainty score of the most confident instances $\mathbf{x}$ from $\text{top}_{\gamma\%}$ of instances in the training set with the highest aleatoric uncertainty:

$$\delta_{\min} = U_E(\beta\%); \ \delta_{\max} = U_A(\gamma\%).$$

Algorithm 1: The HUQ algorithm with MD for epistemic UE and SR for aleatoric UE.

Input : Target text $\mathbf{x}$,
Some dataset $D = \{\mathbf{x}_i\}_{i=1}^N$,
Hyperparameters: $\delta_{\min}$, $\delta_{\max}$, α
Output : Uncertainty score $U_{\text{HUQ}}(\mathbf{x})$

$U_{\text{E}}(\mathbf{x}) \leftarrow \text{MD}(\mathbf{x})$; $U_{\text{A}}(\mathbf{x}) \leftarrow \text{SR}(\mathbf{x})$
$\mathcal{X}_{\text{ID}} \leftarrow \{x \colon U_{\text{E}}(x) \leq \delta_{\min}\}$;
$\mathcal{D}_{\text{ID}} = \{x_i \colon x_i \in \mathcal{D}, \cap x_i \in \mathcal{X}_{\text{ID}}\}$
if $\mathbf{x} \in \mathcal{X}_{ID}$ **then** /*When $\mathbf{x}$ is ID*/
  $\mathcal{X}_{\text{AID}} \leftarrow \{x \colon U_{\text{A}}(x) > \delta_{\max}\}$
  if $\mathbf{x} \in \mathcal{X}_{AID}$ **then**
    $U_{\text{HUQ}}(\mathbf{x}) \leftarrow R(U_{\text{A}}(\mathbf{x}), D)$
  else
    $U_{\text{HUQ}}(\mathbf{x}) \leftarrow R(U_{\text{A}}(\mathbf{x}), D_{\text{ID}})$
  end
else /*When $\mathbf{x}$ is not ID*/
  $U_{\text{HUQ}}(\mathbf{x}) \leftarrow$
  $(1-\alpha)R(U_{\text{E}}(\mathbf{x}), D) + \alpha R(U_{\text{A}}(\mathbf{x}), D)$
end

5 Experimental Setup

5.1 Models

We experiment with two pre-trained Transformers: ELECTRA ("electra-base-discriminator") (Clark et al., 2020) and BERT ("bert-base-uncased") (Devlin et al., 2019) with 110 million parameters. We use a spectral normalization of the weight matrix in the penultimate linear layer of the classification heads of the models (Liu et al., 2020) as it can be helpful for density-based methods (Vazhentsev et al., 2022). The details on the model hyperparameter optimization procedure and optimal values are presented in Appendix A. To report the deviation of results, for each experiment, we train 5 models with optimal hyperparameters, but different random seeds.

5.2 Datasets

There are several tasks that contain highly subjective data, e.g., toxicity detection, particularly detecting implicit hate and sentiment analysis. We conduct experiments on five datasets for toxicity detection: PARADETOX (Logacheva et al., 2022), JIGSAW with binary labels,[2] a collection of tweets with annotation of hate and offensive language (TWITTER; Davidson et al., 2017), TOXIGEN (Hartvigsen et al., 2022), and IMPLICITHATE (ElSherief et al., 2021); and three multi-class classification tasks with high ambiguity: 20 NEWS GROUPS (Lang, 1995), Stanford Sentiment Treebank with 5 classes (SST-5; Socher et al.,

[2] Jigsaw Kaggle Toxic Comment Classification Dataset.

Model	Method	Epistemic	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	27.17±4.95	70.97±6.07	112.12±17.39	887.14±16.89	**380.15±19.74**	433.44±33.44	446.07±4.59	**3529.31±62.46**
	MD	-	11.82±1.79	66.24±6.97	100.99±19.30	912.87±27.59	386.05±52.70	436.26±31.58	458.80±11.72	4692.51±249.35
	HUQ (ours)	MD	**11.27±2.27**	**63.69±5.50**	**95.05±11.22**	**878.34±16.30**	385.99±31.49	**383.24±34.26**	**433.78±4.77**	3550.72±57.03
BERT	SR	-	21.83±5.02	76.36±3.84	72.88±11.20	896.71±10.93	441.35±39.75	342.56±25.88	495.25±21.38	**4050.21±42.37**
	MD	-	10.39±0.97	74.49±4.66	93.96±14.54	932.50±26.00	426.26±47.47	322.68±12.01	**460.70±9.66**	5097.01±335.93
	HUQ (ours)	MD	**9.71±1.37**	**74.33±2.64**	**70.53±9.17**	**896.30±22.73**	**416.24±18.19**	**302.39±23.64**	464.64±11.09	4051.15±68.20

Table 1: AUC-RC↓ results for HUQ-MD and baselines. Best results for each model and dataset shown in bold.

Model	Method	Epistemic	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	27.17±4.95	70.97±6.07	112.12±17.39	887.14±16.89	380.15±19.74	433.44±33.44	446.07±4.59	3529.31±62.46
	DDU	-	15.30±1.36	76.29±6.94	170.74±26.88	915.49±25.77	385.64±60.20	398.23±29.35	448.12±10.68	4711.31±348.28
	HUQ (ours)	DDU	**14.63±3.39**	**63.90±4.78**	**110.12±10.75**	**870.22±11.34**	**379.39±42.36**	**371.43±32.98**	**429.30±5.68**	**3514.49±61.13**
BERT	SR	-	21.83±5.02	76.36±3.84	**72.88±11.20**	**896.71±10.93**	441.35±39.75	342.56±25.88	495.25±21.38	4050.21±42.37
	DDU	-	13.02±2.81	76.31±9.07	223.77±73.40	925.60±30.92	446.28±78.86	305.67±13.42	**462.29±9.04**	4819.17±251.74
	HUQ (ours)	DDU	**11.77±2.18**	**73.72±2.94**	74.47±8.47	903.38±37.43	**426.43±39.46**	**294.45±18.78**	467.16±12.97	**4033.59±36.82**

Table 2: AUC-RC↓ results for HUQ-DDU and baselines. Best reults for each model and dataset shown in bold.

Model	Method	Epistemic	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	27.17±4.95	70.97±6.07	112.12±17.39	**887.14±16.89**	**380.15±19.74**	433.44±33.44	446.07±4.59	**3529.31±62.46**
	RDE	-	9.04±1.88	**63.22±4.55**	93.08±9.05	1065.67±23.22	391.57±29.67	432.03±16.77	451.07±13.44	5759.07±149.02
	HUQ (ours)	RDE	**8.89±1.72**	63.37±4.92	**91.83±10.17**	904.80±27.54	380.58±23.58	**366.45±19.96**	**424.47±7.05**	3532.58±60.23
BERT	SR	-	21.83±5.02	76.36±3.84	72.88±11.20	896.71±10.93	441.35±39.75	342.56±25.88	495.25±21.38	4050.21±42.37
	RDE	-	**8.55±1.83**	72.68±3.47	74.01±10.06	1033.53±23.57	445.15±22.66	331.14±12.94	**470.37±10.42**	6299.17±443.67
	HUQ (ours)	RDE	8.55±1.83	**72.60±2.87**	**68.68±6.03**	**885.65±15.82**	**424.28±22.04**	**289.65±9.81**	476.81±18.02	**4046.09±46.42**

Table 3: AUC-RC↓ results for HUQ-RDE and baselines. Best results for each model and dataset shown in bold.

2013), and AMAZON REVIEWS (McAuley and Leskovec, 2013) (sports and outdoors categories). Note that for TOXIGEN and IMPLICITHATE, implicit hate speech accounts for more than 95% of the positive class. The TWITTER dataset does not contain a predefined test set, so we create it by ourselves. It is constructed from the documents with high annotator disagreement. In all other cases, we use original test sets. See Appendix B for dataset statistics and the analysis of their ambiguity.

To reduce the computational burden of the experiments, the datasets are randomly subsampled. For training, we sample 10% from AMAZON, IMPLICITHATE, and JIGSAW; and 20% from PARADETOX. For evaluation, we sample 10% from PARADETOX, IMPLICITHATE, and JIGSAW.

5.3 Metrics

Selective classification differs from the standard classification task as low certainty predictions are rejected and deferred to alternate procedures, e.g., human review. Therefore, for performance evaluation in this task, a special metric is used: area under the risk coverage curve (AUC-RC; El-Yaniv and Wiener, 2010). Consider all predictions in a dataset are sorted in ascending order by uncertainty, so we can discard some % of the most uncertain predictions. The % of predictions remaining after that is called a coverage rate, and the total loss of the remaining predictions is called the selective risk. The RC curve plots a dependence of the selective risk from the coverage rate. Finally, the AUC-RC is a cumulative sum of the selective losses for each coverage rate. Lower values of AUC-RC indicate better performance.

5.4 Hyperparameter Selection for HUQ

To find optimal hyperparameters for HUQ, we select 20% of the training set as a validation set and optimize AUC-RC on it, using a grid search. For each variant of models trained with different random seeds, we select its specific set of hyperparameters. The hyperparameter grid is the following: $\alpha \in [0; 1]$ with a step size 0.1; $\delta_{\min} \in \{0\%, 0.05\%, 0.1\%, 0.15\%, 0.2\%\}$; $\delta_{\max} \in \{0.9\%, 0.95\%, 1.0\%\}$. The values of $\delta_{\max}$ and $\delta_{\min}$ in % are converted into absolute values, when we apply them to the test data.

6 Results

In our illustrative example of the two moons dataset in Figure 1, the state-of-the-art epistemic UE methods, MD and RDE, separate the ID area from the remaining feature space well. However, the middle area between the two classes is marked with high confidence, yet for SR and Entropy, this area is marked as highly uncertain due to the presence of instances with high aleatoric uncertainty. HUQ, which combines aleatoric and epistemic uncertainty, accurately detects both areas of uncertainty, thereby overcoming the weaknesses of aleatoric and epistemic uncertainty individually applied.

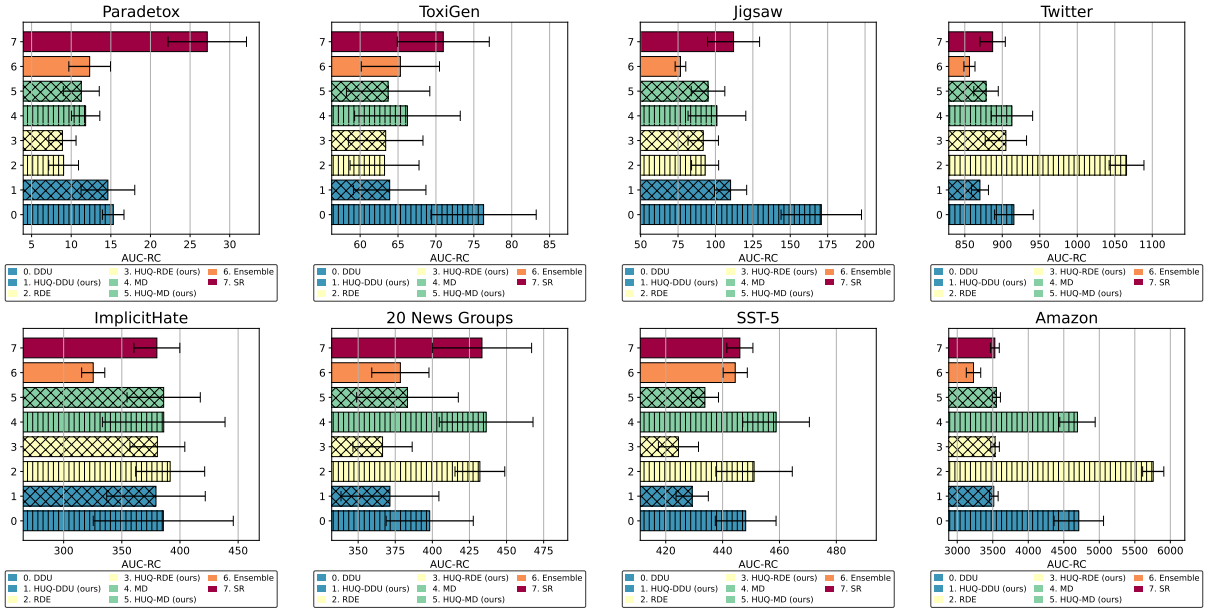

Figure 2: Overall comparison by AUC-RC↓ of various UE methods for ELECTRA.

6.1 HUQ Against its Base Methods

When presenting results, we denote HUQ with a specific base epistemic UE method as HUQ (<UE method>). Note that in the main part of the paper, we present the results only with SR as a base aleatoric UE method. The results for entropy are very similar to SR and are presented in Appendix E.

HUQ-MD yields significant improvements over its base methods (MD and SR) on $6/8$ datasets for both ELECTRA and BERT (see Table 1). The largest improvements are achieved on 20 NEWS GROUPS and PARADETOX, where HUQ reduces AUC-RC by 13.0% and 4.9% (ELECTRA), and 6.6% and 7.0% (BERT).

HUQ-DDU produces improvements over DDU and SR on all 8 datasets for ELECTRA and on 5 datasets for BERT (see Table 2). For ELECTRA, HUQ produces large effects on PARADETOX and TOXIGEN with 4.6% and 11% AUC-RC reduction, and with BERT on PARADETOX with a 10.6% reduction. Interestingly, vanilla DDU is significantly outperformed by SR for ELECTRA on JIGSAW. Applying HUQ addresses this issue, and improves on the results using SR by 1.8%.

The results for HUQ-RDE are more ambiguous than for DDU and SR (see Table 3). RDE is a good method for selective classification and is a hard-to-beat baseline for HUQ. This is because, in addition to OOD detection RDE computes a covariance matrix for each class, thereby making it suitable for identifying decision boundaries. For RDE as the base epistemic UE method, HUQ improves results on 4 datasets for ELECTRA and on 6 datasets for BERT. On some datasets, HUQ does not improve on SR and RDE, e.g., for TWITTER (ELECTRA) and PARADETOX (BERT). However, on others, HUQ shows big improvements in RC-AUC, e.g., 18.0% for 20 NEWS GROUPS (ELECTRA) and 5.0% for TOXIGEN (BERT).

Overall, we see that HUQ usually improves upon its base methods, but in some cases, retains the same performance. We suspect that the configurations where HUQ does not outperform the baselines are due to the presence of large covariate shifts between the training and test data. We discuss this in detail in Section 7.

6.2 Overall Comparison

Here, we compare HUQ in selective classification tasks with various other UE techniques, including strong, yet computationally intensive deep ensembles (DE Lakshminarayanan et al., 2017) and SelectiveNet (Geifman and El-Yaniv, 2019) specifically designed for selective classification, but previously tested only in computer vision. Figure 2 presents results for the ELECTRA model and Figure 9 in Appendix D presents results for BERT.

The base epistemic UE methods sometimes cannot outperform even the weak SR baseline or even fall behind it by a large margin. It is especially noticeable for RDE on TWITTER and AMAZON reviews and for DDU on JIGSAW and AMAZON re-

views. This effect might appear because the majority of model mistakes arise from ambiguity rather than OOD instances, while these methods are better suitable for OOD detection. On some datasets, it is very hard to overcome the weak SR baseline. For example, on IMPLICITHATE and AMAZON, only DE confidently outperforms SR.

The results for our implementation of SelectiveNet for text classification models and the detailed experimental setup for this method are presented in Appendix F. On all considered datasets, SelectiveNet never outperforms the SR baseline and significantly falls behind it.

Variants of HUQ are usually the best or the second best after DE. For example, HUQ outperforms this strong baseline on PARADETOX, 20 NEWS GROUPS, and SST-5. However, while DE introduces computational overhead of 400%, HUQ requires additionally less than 5% of standard model inference time (see Table 15 in Appendix G).

6.3 Analyses

Hyperparameter for mixing aleatoric and epistemic uncertainty scores in HUQ. When varying the hyperparameter α, we change the impact of aleatoric and epistemic uncertainty for the final score. Figure 3 reports the impact of α on the TOXIGEN dataset. When α is close to 0, the performance of the total score approximates the epistemic uncertainty represented by MD, which is even worse in terms of AUC-RC than the SR baseline. When α is close to 1, we use solely the SR score in the mixture of uncertainties, while treating AID, ID, and other instances differently, which results in better performance compared to vanilla SR. The best results are obtained when we select α on the validation set. We can see that obtained $\alpha = 0.5$ is very close to its optimum on the test set. HUQ-MD in this case outperforms MD by 10.6% and SR by 9.6% in terms of AUC-RC. This again illustrates the importance of mixing different types of uncertainties for selective classification. Similar charts for other considered datasets are presented in Figure 6 in Appendix C and for other hyperparameters in Figures. 7 and 8 in Appendix C.

Qualitative analysis. Table 16 in Appendix H presents several instances from various datasets, as well as model predictions and their normalized uncertainty scores. The qualitative analysis reveals that baseline uncertainty scores MD and SR may be high regardless of whether a classification of

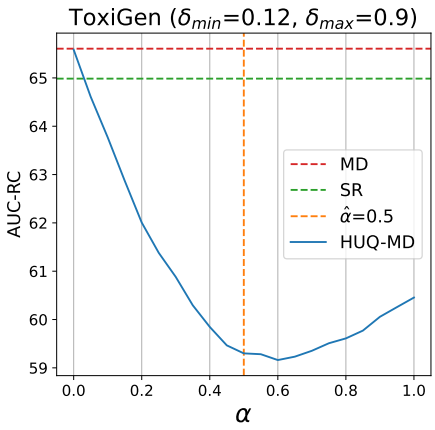

Figure 3: AUC-RC↓ for different values of α in HUQ on TOXIGEN using ELECTRA. The vertical line denotes $\hat{\alpha}$ selected on the validation set.

an instance is correct. For example, we see that four correctly classified instances in PARADETOX are marked with high uncertainty by at least one of the methods. Moreover, MD and SR disagree with each other: MD yields high uncertainty scores for the first two instances, whereas SR produces low uncertainty. For the last two instances, the pattern is reversed. In all of these cases, the MD score is not low enough to consider instances as ID. Therefore, HUQ-MD linearly mixes the SR and MD scores, producing more balanced results with moderately low uncertainty, which is consistent with the fact that classifications are correct.

For the last example from Jigsaw, MD falls below a threshold α obtained for this dataset. Consequently, the example is classified as an ID instance, leading to the HUQ-MD score being equal to the SR score for this particular case. Contrary to MD, which yields low uncertainty, high uncertainty of SR and HUQ correctly indicates a prediction error.

For two examples, HUQ-MD contradict the results. Specifically, in the third example of ToxiGen and the second example of Jigsaw, the predictions are accurate, but uncertainty is moderately high. This discrepancy arises from both SR and MD being erroneously high. In such cases, the hybrid method is unable to correct the uncertainty score.

7 Limitations

While HUQ outperforms individual aleatoric and epistemic UE methods for most datasets considered, for some, the effects are negligible. To understand this pattern, we analyze the difference between the training and test sets. We generate latent representations of instances in the datasets

	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
F1-score	0.02	0.0	0.69	0.35	0.66	0.08	0.0	0.75
Mean Impr. HUQ-DDU, %	4.60	11.0	1.80	1.90	0.02	7.20	3.9	0.30

Table 4: The performance of separation instances into train and test datasets. The classification is performed by the logistic regression model trained on latent feature representations obtained from ELECTRA.

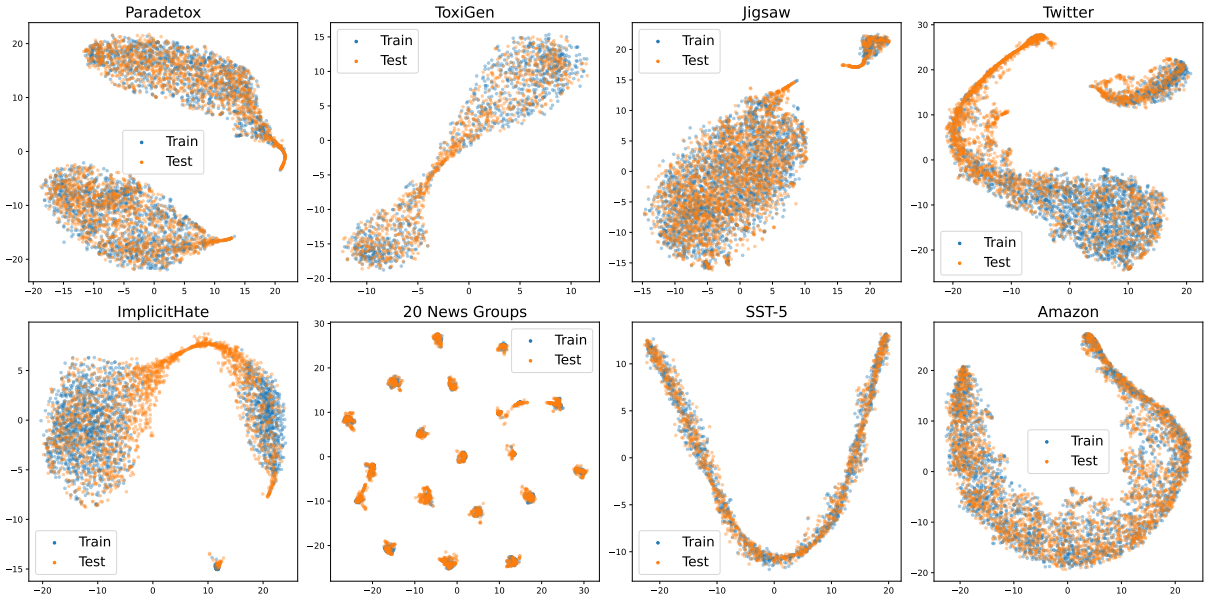

Figure 4: Visualization of the t-SNE decomposition of latent representations obtained from the fine-tuned ELECTRA model for the train and test datasets.

using a fine-tuned ELECTRA model and fit a logistic regression model to discriminate between train and test sets using these representations as features. Good performance of the discriminator indicates a covariance shift between the training and test data, while bad performance indicates that instances come from the same distribution.

Table 4 presents F1 scores for this task aligned with the performance gains of HUQ-DDU in percentages over the best method from the pair <SR, DDU>. As we can see, high F1 scores often correspond to low values of performance gains (the Spearman rank correlation $= 0.8$). This means that HUQ is unlikely to provide improvements to the base methods for the tasks with big covariate shifts. In our analysis, this is due to prediction mistakes primarily arising from OOD instances, which are well-handled by epistemic UE methods.

Visualizing the differences between the datasets using a t-SNE decomposition of the latent representations (see Figure 4), we can see that for IMPLICITHATE and TWITTER, where HUQ does not provide improvements, some regions of the test data are not covered by the training set. For PARADETOX and TOXIGEN, on the other hand, the training dataset completely overlays all regions of the test data, and using HUQ improves AUC-RC on the base methods.

8 Conclusion

In this work, we proposed a hybrid uncertainty quantification method for selective text classification. It combines pre-existing methods for aleatoric and epistemic uncertainty, providing scores of total uncertainty. Experimentally, we find that HUQ usually outperforms in terms of RC-AUC other UE methods that aim at quantifying only one type of uncertainty. In real terms, the improved uncertainty estimation offered by our method affords improved identification of erroneous predictions for ambiguous text classification tasks.

Although the HUQ method often provides better results, there are some cases where it is unable to surpass its base methods and performs at a comparable level to them. In our analysis of these examples, we find that this issue arises when there is a substantial covariate shift between the training and test data. In future work, we are planning to analyze other factors that affect the performance of UE methods in selective classification tasks. Our goal is to achieve more consistent and stable improvements over baselines across diverse datasets.

Acknowledgements

We are very grateful to Zeerak Talat for generously sharing their expertise in toxicity detection, offering valuable suggestions for text edits, and the help with the work in general. We thank anonymous reviewers for their insightful feedback towards improving this paper. The financial support was provided by the Russian Science Foundation, grant 20-71-10135.

Ethical Considerations

The task of uncertainty estimation is one that is closely tied to the construction of ethical machine learning methods, as it pertains to the identification of potential misclassified instances. For the task of toxic content classification, uncertainty estimation is particularly important due to the speech concerns surrounding toxicity detection. Moreover, toxicity detection has shown disparate performance along gendered and racialized lines, uncertainty estimation provides an avenue for identifying when a model may no longer be applied without further improvement. However, while uncertainty estimation may have potential benefits to the tasks under the umbrella of abusive language detection, approaching misclassifications and uncertainty without an intersectional (Crenshaw, 1991) lens, and without appropriate measures for deep engagements with affected communities may propagate issues of social control, and particularly of enforcing respectability politics of language use. It is therefore important to understand that uncertainty estimation can only provide a partial perspective to the challenges that are faced in abusive language detection. For instance, data that is mislabeled, or labeled such that it propagates stereotypes can exhibit low levels of uncertainty while being undesirable in relation to the goal of equitable machine learning methods for content moderation.

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1192–1201. PMLR.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

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.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph.D. thesis, University of Cambridge.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS 2017, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.

Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97

of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and Chang-Tien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8362–8372. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Aleksandr Petiushko, Artem Shelmanov, Artem Vazhetsev, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS 2017, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 331–339. Morgan Kaufmann.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, volume 31, pages 7167–7177.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512. Curran Associates, Inc.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Andrey Malinin and Mark J. F. Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7047–7058.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.

Sarah T. Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Peter J Rousseeuw. 1984. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks.

In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13888–13899.

Nanna Thylstrup and Zeerak Waseem. 2020. Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. *SSRN Electronic Journal*.

Joost R. van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language 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 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

A Hyperparameter Values and Hardware Configuration

Model	Dataset	Accuracy Score	Learning Rate	Num. Epochs	Batch Size	Weight Decay
ELECTRA	PARADETOX	0.972	2e-05	10	8	0.10
	TOXIGEN	0.858	5e-05	3	32	0.00
	JIGSAW	0.967	3e-05	13	64	0.00
	TWITTER	0.976	9e-06	4	16	0.01
	IMPLICITHATE	0.707	7e-05	15	64	0.01
	20 NEWS GROUPS	0.897	3e-05	12	64	0.00
	SST-5	0.585	9e-06	4	16	0.01
	AMAZON	0.736	5e-05	2	32	0.00
BERT	PARADETOX	0.971	3e-05	7	16	0.10
	TOXIGEN	0.845	5e-05	2	32	0.00
	JIGSAW	0.964	5e-05	3	32	0.00
	TWITTER	0.978	7e-06	6	32	0.00
	IMPLICITHATE	0.702	7e-05	9	64	0.01
	20 NEWS GROUPS	0.909	7e-05	9	64	0.01
	SST-5	0.533	7e-05	15	64	0.00
	AMAZON	0.705	9e-06	3	16	0.01

Table 5: Optimal hyperparameters for each model and dataset.

The optimal hyperparameters are obtained using Bayesian optimization with early stopping. We train a model on 80% of the training dataset and validate on the remaining 20%. The optimal hyperparameters are selected according to the best accuracy score on the validation set. After the hyperparameters are selected, we use them to fine-tune the model on the full training set. The hyperparameter grid is the following:

Learning rate: [5e-6, 6e-6, 7e-6, 9e-6, 1e-5, 2e-5, 3e-5, 5e-5, 7e-5, 1e-4];
Num. of epochs: $\{n \in \mathbb{N} \mid 2 \leq n \leq 15\}$;
Batch size: [4, 8, 16, 32, 64];
Weight decay: [0, 1e-2, 1e-1].

Table 6 presents the hardware configuration used in experiments. In addition, we provide the approximate number of GPU hours that are needed for training and evaluating all models for all datasets.

CPU	2 Intel Xeon Platinum 8168, 2.7 GHz
CPU Cores	24
GPU	NVIDIA Tesla v100 GPU
GPU Memory	32 GB
GPU Hours	272

Table 6: Hardware configuration used in this work and the approximate number of GPU hours spent for running experiments.

B Dataset Statistics and Analysis of Ambiguity in Datasets

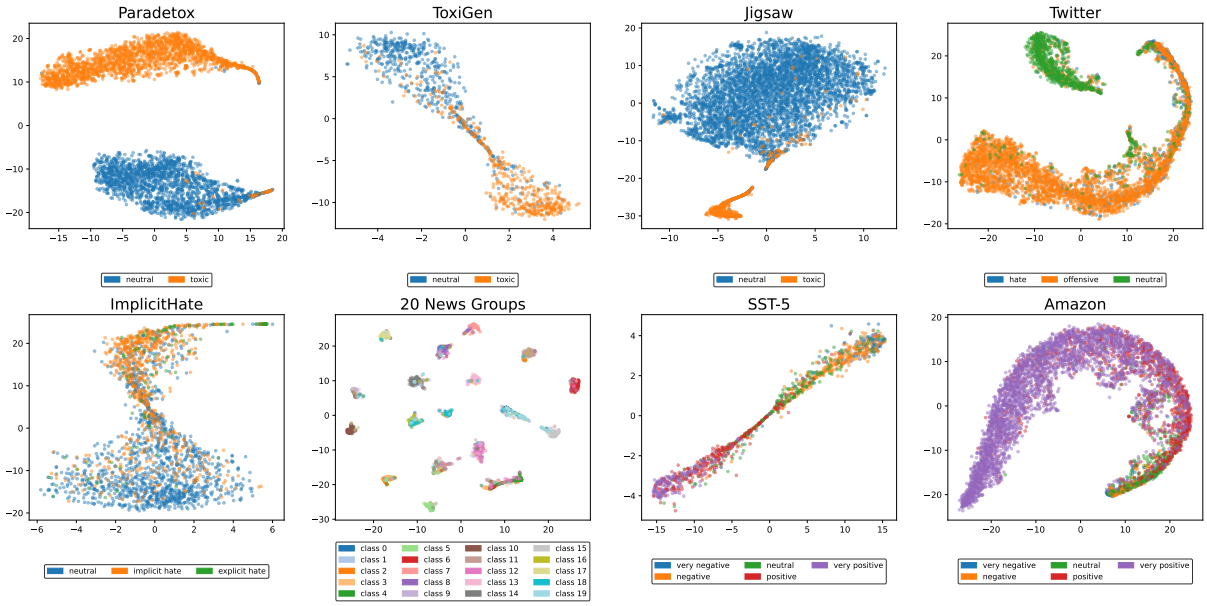

Figure 5: Visualization of the t-SNE decomposition of latent representations from the fine-tuned ELECTRA model for various test sets. Different colors indicate different classes.

Figure 5 presents the t-SNE decomposition of latent representations from ELECTRA for various test sets, where classes are marked with different colors. For TOXIGEN, JIGSAW, and IMPLICITHATE, we can see that there is no clear boundary between the "neutral" and "toxic" classes. For SST-5 and AMAZON, we can see a smooth transition from the "very negative" to the "very positive" classes. This illustration reveals the presence of noisy and ambiguous instances in these datasets.

Table 7 presents the dataset statistics with the number of instances in the test and training sets and the number of labels.

	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
Train	39.5K	9.0K	159.6K	24.8K	21.5K	11.3K	8.5K	207.4K
Test	-	0.9K	-	-	-	7.5K	1.1K	29.6K
# Labels	2	2	2	3	3	20	5	5

Table 7: Dataset statistics. For SST-5, we used the validation set as the test set. For datasets, where the test data is not given we split the entire training dataset into the training and test parts as described in Section 5.2.

C Contribution of Different Components of HUQ

Figures. 6 to 8 present the dependence of the AUC-RC score when varying one of the hyperparameters in HUQ, while others are fixed to optimal values. According to the results, the most valuable hyperparameter of HUQ is α. In Figure 6, for all datasets, except IMPLICITHATE, we see that there exists an optimal value of α different from 0 or 1 that gives the smallest AUC-RC. This means that for these datasets, the contributions of both types of uncertainties are important. In addition, we can see that our validation strategy finds $\hat{\alpha}$ close to its optimal value.

Hyperparameters $\delta_{\min}$ and $\delta_{\max}$ contribute to the final score, but their effect is less significant. Nevertheless, it is crucial to take into account all components of HUQ to achieve the best results.

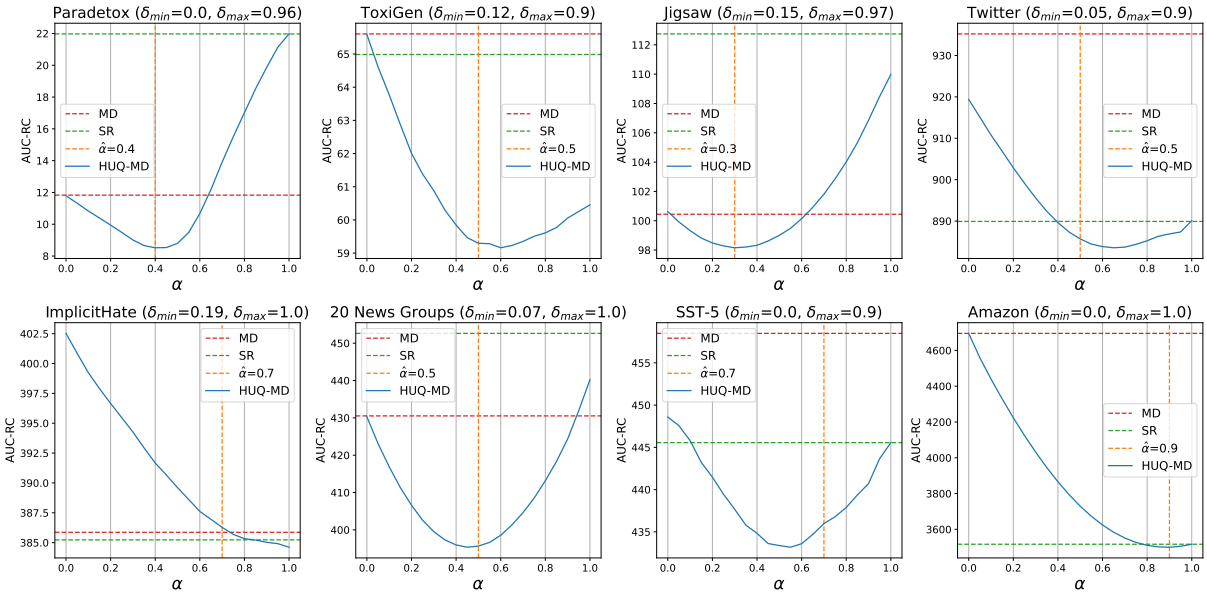

Figure 6: Dependence of AUC-RC↓ when varying the parameter α for ELECTRA. The parameters $\delta_{\min}$ and $\delta_{\max}$ are fixed and presented in the title. The vertical line indicates the selected optimal value $\hat{\alpha}$ on the validation set.

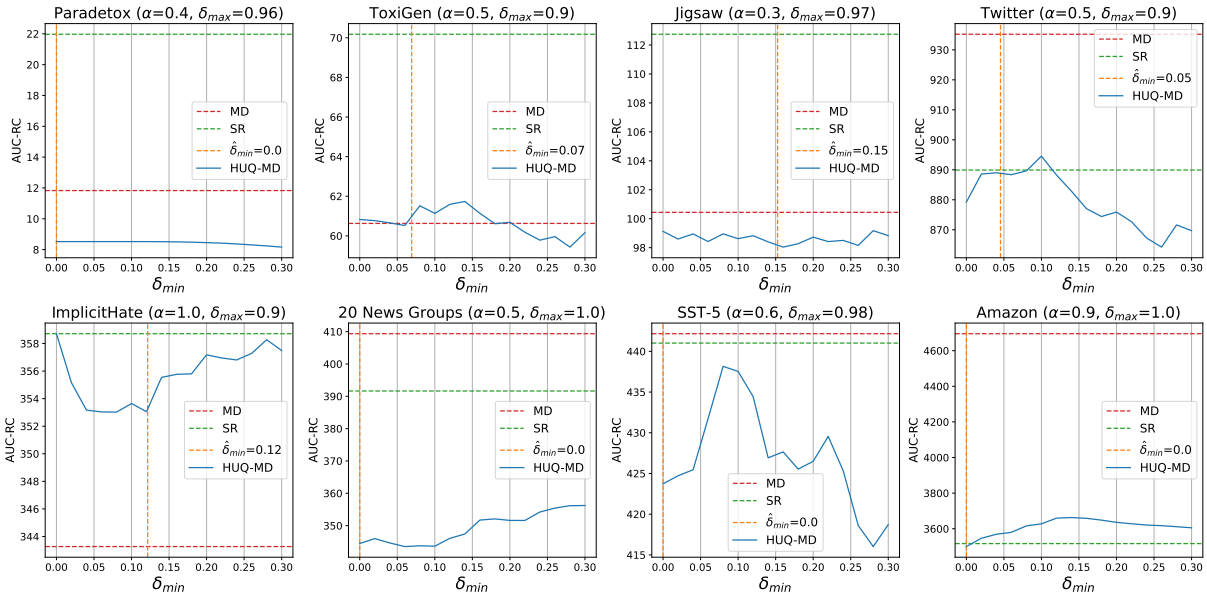

Figure 7: Dependence of AUC-RC↓ when varying the parameter $\delta_{\min}$ for ELECTRA. The parameters α and $\delta_{\max}$ are fixed and presented in the title. The vertical line indicates the selected optimal value $\hat{\delta}_{\min}$ on the validation set.

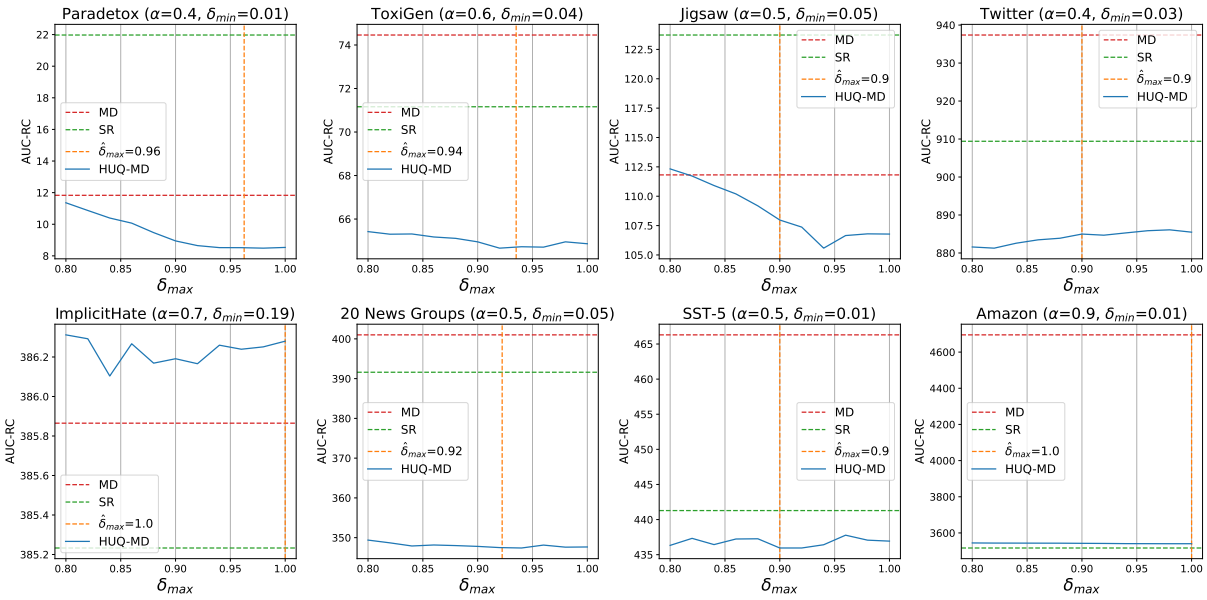

Figure 8: Dependence of AUC-RC↓ when varying the parameter $\delta_{\max}$ for ELECTRA. The parameters α and $\delta_{\min}$ are fixed and presented in the title. The vertical line indicates the selected optimal value $\hat{\delta}_{\max}$ on the validation set.

D Overall Comparison of UE Methods for BERT

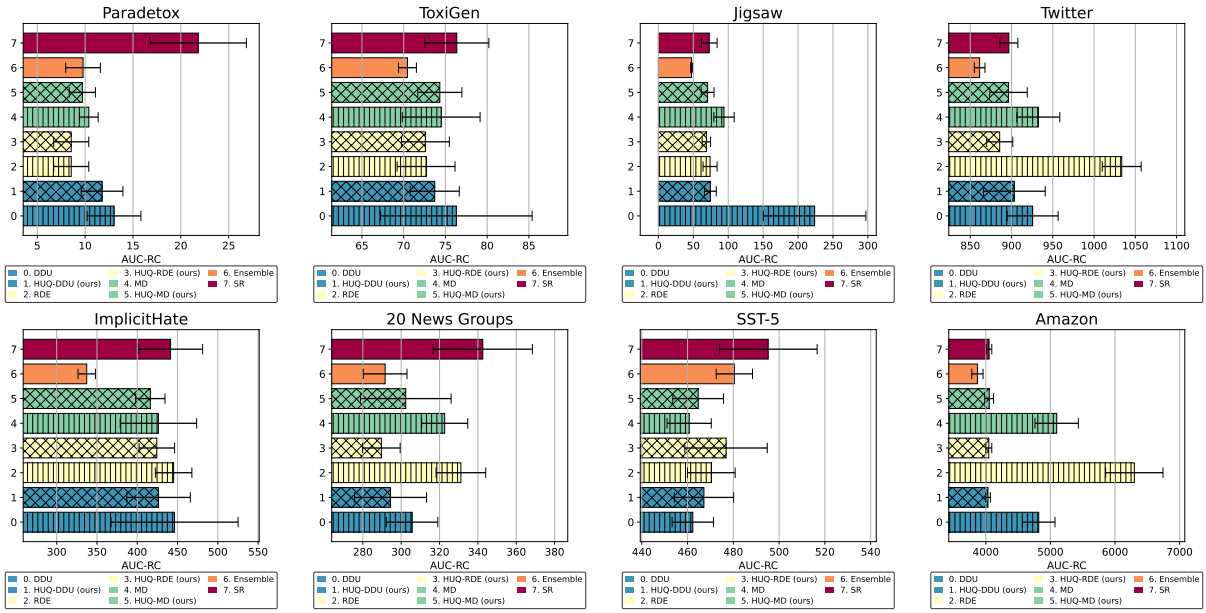

Figure 9: Overall comparison by AUC-RC↓ of UE methods for the BERT model.

E Additional Experiments with Different Aleatoric Uncertainty Estimation Methods

Tables 8 to 10 present the comparison of AUC-RC when using SR or Entropy as measures of aleatoric uncertainty in various versions of HUQ. Only for SST-5, we see a small significant difference between them: SR is better than Entropy in terms of AUC-RC by 2.5% in HUQ-MD, by 2.7% in HUQ-DDU, and by 1.7% in HUQ-RDE.

Model	Method	Epistemic	Aleatoric	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	-	27.17±4.95	70.97±6.07	112.12±17.39	887.14±16.89	**380.15±19.74**	433.44±33.44	446.07±4.59	3529.31±62.46
	Entropy	-	-	27.23±5.10	70.97±6.07	112.09±17.31	887.32±17.18	380.23±20.25	435.01±33.60	448.66±3.53	**3521.78±71.23**
	HUQ	MD	SR	**11.27±2.27**	63.69±5.50	**95.05±11.22**	**878.34±16.30**	385.99±31.49	**383.24±34.26**	**433.78±4.77**	3550.72±57.03
	HUQ (ours)	MD	Entropy	11.29±2.25	**63.68±5.50**	95.08±11.20	879.04±16.83	387.19±36.12	383.76±34.04	444.49±3.34	3545.13±60.40
BERT	SR	-	-	21.83±5.02	76.36±3.84	72.88±11.20	896.71±10.93	441.35±39.75	342.56±25.88	495.25±21.38	**4050.21±42.37**
	Entropy	-	-	21.82±4.99	76.36±3.84	72.88±11.20	902.59±20.84	437.93±44.28	345.43±23.05	496.21±22.07	4078.42±51.21
	HUQ	MD	SR	**9.71±1.37**	74.33±2.64	**70.53±9.17**	**896.30±22.73**	416.24±18.19	302.39±23.64	**464.64±11.09**	4051.15±68.20
	HUQ (ours)	MD	Entropy	9.72±1.37	**74.31±2.62**	70.57±9.12	899.78±23.24	**413.77±20.26**	**301.99±23.81**	464.90±11.17	4082.83±76.64

Table 8: The comparison of SR and Entropy as measures of aleatoric uncertainty in HUQ-MD for ELECTRA and BERT models. The best results for each model are shown in bold.

Model	Method	Epistemic	Aleatoric	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	-	27.17±4.95	70.97±6.07	112.12±17.39	887.14±16.89	380.15±19.74	433.44±33.44	446.07±4.59	3529.31±62.46
	Entropy	-	-	27.23±5.10	70.97±6.07	112.09±17.31	887.32±17.18	380.23±20.25	435.01±33.60	448.66±3.53	3521.78±71.23
	HUQ	DDU	SR	**14.63±3.39**	**63.90±4.78**	**110.12±10.75**	**870.22±11.34**	379.39±42.36	**371.43±32.98**	**429.30±5.68**	3514.49±61.13
	HUQ (ours)	DDU	Entropy	14.79±3.05	63.92±4.78	113.12±16.99	872.40±13.55	**378.47±42.68**	371.92±32.89	441.20±4.11	**3512.43±68.07**
BERT	SR	-	-	21.83±5.02	76.36±3.84	**72.88±11.20**	**896.71±10.93**	441.35±39.75	342.56±25.88	495.25±21.38	4050.21±42.37
	Entropy	-	-	21.82±4.99	76.36±3.84	72.88±11.20	902.59±20.84	437.93±44.28	345.43±23.05	496.21±22.07	4078.42±51.21
	HUQ	DDU	SR	**11.77±2.18**	**73.72±2.94**	74.47±8.47	903.38±37.43	426.43±39.46	294.45±18.78	467.16±12.97	**4033.59±36.82**
	HUQ (ours)	DDU	Entropy	11.77±2.18	74.00±3.42	74.47±8.46	905.05±35.51	**424.77±42.83**	**294.32±18.85**	**466.94±12.89**	4066.18±47.16

Table 9: The comparison of SR and Entropy as measures of aleatoric uncertainty in HUQ-DDU for ELECTRA and BERT models. The best results for each model are shown in bold.

Model	Method	Epistemic	Aleatoric	PARADETOX	TOXIGEN	JIGSAW	TWITTER	IMPLICITHATE	20 NEWS GROUPS	SST-5	AMAZON
ELECTRA	SR	-	-	27.17±4.95	70.97±6.07	112.12±17.39	**887.14±16.89**	**380.15±19.74**	433.44±33.44	446.07±4.59	3529.31±62.46
	Entropy	-	-	27.23±5.10	70.97±6.07	112.09±17.31	887.32±17.18	380.23±20.25	435.01±33.60	448.66±3.53	3521.78±71.23
	HUQ	RDE	SR	**8.89±1.72**	**63.37±4.92**	**91.83±10.17**	904.80±27.54	380.58±23.58	**366.45±19.96**	**424.47±7.05**	3532.58±60.23
	HUQ (ours)	RDE	Entropy	8.89±1.72	63.37±4.93	91.84±10.17	898.43±18.71	380.57±23.72	366.77±20.64	431.11±6.18	**3515.40±67.87**
BERT	SR	-	-	21.83±5.02	76.36±3.84	72.88±11.20	896.71±10.93	441.35±39.75	342.56±25.88	495.25±21.38	4050.21±42.37
	Entropy	-	-	21.82±4.99	76.36±3.84	72.88±11.20	902.59±20.84	437.93±44.28	345.43±23.05	496.21±22.07	4078.42±51.21
	HUQ	RDE	SR	**8.55±1.83**	72.60±2.87	**68.68±6.03**	**885.65±15.82**	424.28±22.04	289.65±9.81	**476.81±18.02**	**4046.09±46.42**
	HUQ (ours)	RDE	Entropy	8.55±1.83	**72.58±2.87**	68.68±6.04	888.11±18.76	**421.93±24.35**	**289.22±10.56**	477.47±18.58	4072.77±54.12

Table 10: The comparison of SR and Entropy as measures of aleatoric uncertainty in HUQ-RDE for ELECTRA and BERT models. The best results for each model are shown in bold.

F Additional Experiments with SelectiveNet

Tables 11 to 13 present the comparison of the SelectiveNet performance (Geifman and El-Yaniv, 2019) with the performance of the SR baseline. The experiments are conducted with the ELECTRA model on the PARADETOX, TOXIGEN, and 20 NEWS GROUPS datasets. SelectiveNet is designed only for a specific coverage, which is fixed during training. Therefore, we select multiple coverage values and for each value, we fine-tune a separate model, following the standard approach for training SelectiveNet. Since the coverage for each model is fixed, the AUC-RC metric is not appropriate for evaluation of this method. Therefore, instead, we use the selective risk for the specified coverage as an evaluation metric. The results show that for the considered text classification datasets, SelectiveNet significantly falls behind the standard SR baseline, which is different from the results obtained by Geifman and El-Yaniv (2019) on computer vision tasks. The optimal hyperparameters for SelectiveNet are presented in Table 14.

Method \ Coverage	0.7	0.8	0.85	0.9	0.95
SR	**7.67±3.56**	**11.17±4.02**	**15.00±4.86**	**22.83±3.43**	**42.83±3.43**
SelectiveNet | SR	16.00±7.80	18.83±14.25	32.17±28.27	32.67±22.31	51.33±28.03
SelectiveNet	12.17±10.91	18.50±12.94	60.50±26.60	44.33±25.94	86.67±22.18

Table 11: Selective risk for various coverages on the PARADETOX dataset. We compare the score from the selective head of the SelectiveNet model with the SR of the SelectiveNet model and SR of the standard ELECTRA model.

Method \ Coverage	0.7	0.8	0.85	0.9	0.95
SR	**59.50±6.28**	**86.00±6.72**	**100.17±5.27**	**115.33±5.32**	**134.50±4.23**
SelectiveNet | SR	75.67±7.81	108.17±6.43	107.83±6.71	138.50±26.33	148.00±10.04
SelectiveNet	101.33±25.31	134.33±5.54	112.67±9.95	158.50±21.80	148.67±10.13

Table 12: Selective risk for various coverages on the TOXIGEN dataset. We compare the score from the selective head of the SelectiveNet model with the SR of the SelectiveNet model and SR of the standard ELECTRA model.

Method \ Coverage	0.7	0.8	0.85	0.9	0.95
SR	**329.00±17.63**	**486.17±23.74**	**617.67±25.15**	**800.83±25.81**	**1012.83±31.40**
SelectiveNet | SR	449.67±44.20	668.00±109.26	1656.33±2062.30	2024.83±2177.35	1070.00±39.01
SelectiveNet	472.17±48.84	794.17±106.07	1759.33±1959.53	2146.50±2108.52	1175.00±47.12

Table 13: Selective risk for various coverages on the 20 NEWS GROUPS dataset. We compare the score from the selective head of the SelectiveNet model with the SR of the SelectiveNet model and SR of the standard ELECTRA model.

Dataset	Coverage	Objective Score	Learning Rate	Num. Epochs	Batch Size	Weight Decay	Reg. Lambda
PARADETOX	0.70	0.98	2e-5	7	4	0.00	30
	0.80	0.98	5e-5	8	32	0.10	40
	0.85	0.98	2e-5	11	8	0.01	1
	0.90	0.98	7e-5	5	64	0.01	10
	0.95	0.98	2e-5	6	64	0.10	32
TOXIGEN	0.70	0.86	2e-5	5	4	0.00	30
	0.80	0.85	1e-5	12	8	0.00	10
	0.85	0.85	3e-5	13	32	0.01	32
	0.90	0.86	2e-5	7	4	0.00	30
	0.95	0.86	3e-5	12	4	0.00	10
20 NEWS GROUPS	0.70	0.88	5e-5	11	32	0.10	10
	0.80	0.89	5e-5	12	8	0.10	40
	0.85	0.89	5e-5	12	8	0.10	40
	0.90	0.88	5e-5	12	8	0.10	40
	0.95	0.87	5e-5	8	32	0.10	40

Table 14: Optimal hyperparameters for the SelectiveNet model. The hyperparameter grid for Reg. Lambda is: [1, 10, 20, 30, 32, 40]. For other hyperparameters, we use the same grids as for the standard model.

G Computation Overhead for Uncertainty Estimation

Table 15 presents the computation time for various UE methods. The HUQ-MD during the inference stage introduces only 0.02% of overhead in comparison with the MD and less than 5% of overhead in comparison with the SR baseline. On the contrary, a deep ensemble of 5 models introduces 400% of overhead in comparison with the MD and the SR, which makes it impractical.

Method	PARADETOX, sec	TOXIGEN, sec	JIGSAW, sec	TWITTER, sec	IMPLICITHATE, sec	20 NEWS GROUPS, sec	SST-5, sec	AMAZON, sec
SR	10.6+0.1	2.9+0.0	43.2+0.3	18.4+0.2	6.9+0.1	18.6+0.1	2.9+0.0	77.8+5.9
MD	11.0+0.1	3.0+0.1	45.2+0.7	19.3+0.3	7.4+0.2	22.1+0.9	3.1+0.0	82.3+6.0
HUQ-MD (ours)	11.0+0.1	3.0+0.1	45.2+0.7	19.3+0.3	7.4+0.2	22.1+0.9	3.1+0.0	82.3+6.0
Deep Ensemble	52.9+0.1	14.3+0.1	216.1+0.6	92.0+0.3	34.4+0.1	92.9+0.1	14.5+0.0	388.8+14.1

Table 15: The computation time for the inference stage for the selected methods.

H Qualitative Analysis

Dataset	Text	Prediction	True Label	SR	MD	HUQ-MD	δ_{min}
PARADETOX	i kinda feel bad saying this but , " yes ! "	Neutral	Neutral	0.02	0.62	0.27	0.0
	He was really scared	Neutral	Neutral	0.01	0.62	0.25	0.0
	that 's how we roll , alt as f*ck	Toxic	Toxic	0.9	0.16	0.45	0.0
	f*ck it , i 'm voting greens .	Toxic	Toxic	0.9	0.14	0.43	0.0
TOXIGEN	the only reason they are so successful is because they work harder than us	Toxic	Neutral	0.85	0.42	0.69	0.0
	what 's a mexicans favourite sport ? cross country	Neutral	Toxic	0.61	0.89	0.71	0.0
	in the end, it is the american taxpayer who pays for all of this	Neutral	Neutral	0.53	0.91	0.66	0.0
JIGSAW	"Regarding text formatting, I'm talking about this version of Michael Di Biase, which you reverted to more than once. And who's this ""we"" in the final line of your comment? "	Neutral	Neutral	0.04	0.37	0.3	0.15
	Right, it's done. I've moved the appropriate comments from this talk page to that one.	Neutral	Neutral	0.80	0.48	0.65	0.2
	You are retarded Learn to type like you're not five years old, MORON!	Toxic	Neutral	0.93	0.16	0.93	0.2
TWITTER	“@SimplyHype: @kobebryant shut the f*ck up, you f*cking trash n*gga, work on another ring” <URL>	Offensive	Toxic	0.77	0.54	0.65	0.05
	Im soooooo tired of this d*ck dyke showing up on my fb feed. Real studs dont fucc n*ggas. Just be bisexual.	Offensive	Toxic	0.77	0.52	0.64	0.05
	Ebola sounds like a hood hoe	Offensive	Toxic	0.30	0.74	0.66	0.0
	In two months tho h*e	Offensive	Offensive	0.15	0.68	0.43	0.05

Table 16: Textual examples from various datasets with uncertainty scores from HUQ-MD for the ELECTRA model. Uncertainty for each method is presented in the range [0-1]. The value indicates percentages of instances in the test dataset with a lower uncertainty score. The higher saturated color indicates higher uncertainty.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No, we used already preprocessed datasets

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.