TECHNISCHE
UNIVERSITÄT
DARMSTADT

EMPIRICAL EVALUATION OF TOTAL, EPISTEMIC AND ALEATORIC UNCERTAINTY OF A BERT-BASED CLASSIFIER FOR HATE SPEECH

Submitted to the Department of Computer Science
of the Technical University of Darmstadt

MASTERTHESIS (30 CP)

for the degree of
Master of Science (M. Sc.)
in the Master's programme *IT Sicherheit*

by

Frank Walter, B. Sc.
Matr.-Nr. 2670522
`frank.walter@stud.tu-darmstadt.de`

Examiner:	Prof. Dr. Dr. Christian Reuter
Supervisor:	Markus Bayer, M. Sc.
Submitted:	January 31, 2024

Wissenschaft und
Technik für Frieden
und Sicherheit

ABSTRACT

Machine learning models for classification generally just output their prediction, without any further explanation. Since these models are never provably correct, there arises the need to supplement this output with additional information to understand better how and why this output was generated. In this thesis, we take a look at uncertainty, particularly two different kinds of uncertainties: aleatoric and epistemic. We propose a novel way for each of these two kinds of uncertainties to be evaluated in a NLP classification task using a BERT hate speech classifier and applying these evaluation methods to assess how well Shannon entropy can be used to calculate these uncertainties through expected entropy and mutual information. Our results mirror other theoretical and practical work that shows that expected entropy is an excellent measure of aleatoric uncertainty, with our mutual information measure barely correlating with epistemic uncertainty. They validate our evaluation method for aleatoric uncertainty, while remaining inconclusive for our evaluation method for epistemic uncertainty as new tools to gauge how well a given metric actually measures the type of uncertainty it is supposed to measure.

ACKNOWLEDGEMENTS

With loving thanks to my family and friends who have supported me in a myriad of ways and without whom this work would not have been possible.

Many thanks to my supervisor Markus Bayer, whose patience and insightful feedback have always been invaluable.

The author gratefully acknowledges the computing time provided to them on the high-performance computer Lichtenberg at the NHR Centers NHR4CES at TU Darmstadt. This is funded by the Federal Ministry of Education and Research, and the state governments participating on the basis of the resolutions of the GWK for national high-performance computing at universities.

CONTENTS

1 INTRODUCTION

1.1 *Motivation*

Since the early 2010s there has been tremendous progress in the field of machine learning (ML). The advent of the Internet with its data collecting capabilities, innovation in the training process *e.g.* AlexNet (Krizhevsky et al., 2017) as well as increasing computing power (Moore, 1965) allowed the jump to new deep learning (DL) models with unprecedented performance in a variety of tasks.

One such field is natural language processing (NLP). From machine translation to chatbots, NLP has been revolutionized by transformer-based models (Vaswani et al., 2017) such as GPT (Radford et al., 2018). One of the main causes for the increasing amount of data available to NLP researchers has been the advent of social media. However, this development has not been without negative consequences. With the proliferation of social media, the prevalence of online harassment and hate speech has exploded (Brand, 2020; Zakrzewski et al., 2021), impacting the mental health especially of the youth (Feiner & Bursztynsky, 2021), damaged discourse *e.g.* through the promotion of anger-inducing content, often in the form of misinformation or toxicity (Merrill & Oremus, 2021) and constituting a major threat to democracy (Nadeem, 2020).

Hate speech can be seen as a cause as well as a symptom of a divided society. Its prevalence constitutes two warning signs at the same time: on the one hand it signals that there is an atmosphere of callousness or even malintent towards the targeted groups up to the point that hate crimes may occur, but on the other hand it also demonstrates there are social circles of hate speech perpetrators and audiences who likely don't support an equal and egalitarian society. Therefore, hate speech classification can play a crucial role in genocide prevention, particularly for areas with a known history of violence against marginalized groups, *e.g.* if combined with geolocalization. An increase in hate speech can be a warning signal for brewing future violence (Quinn, 2013), since genocide is best understood as a process and social phenomenon rather than an event (Rosenberg, 2012). If we look at the ten stage model of genocide, we find that hate speech lies at the core of three of the first four stages, as well as the last one.[1]

How to deal with these problems? While automated classification can fail spectacularly (Vincent, 2018) and is vulnerable to adversarial attacks (Menn, 2015; Su et al., 2019), its use is ubiquitous and considering the sheer amount of data shared on social media a purely human-based classification process is highly expensive. Purely automated classification, with a problem case such as harassment in particular, suffers from the problem being hard to specify formally given that it can occur through a wide variety of ways. As with many machine learning problems, we, therefore, have a problem underspecification at hand (Krakovna et al., 2020). Additionally, we often have more parameters in

[1] Stanton, G. (2020), *The Ten Stages of Genocide*. https://archive.ph/FQK5A

our DL models than we have training data points, which leads to our model itself also being underspecified. D'Amour et al. (2020) claim that the latter leads to our models likely being brittle for samples out of our training distribution, as many near-optimal *IID* predictors could be chosen in the training set, without encoding the proper underlying structure. This phenomenon is also called mismatched objectives (Doshi-Velez & Kim, 2017) and is distinct from distributional drift, as the latter describes a situation in which there are structural differences between training and deployment data, requiring the predictor to do more generalization.

Since we generate our models from our training data inductively, they cannot be provably correct, making them and their predictions inherently uncertain (Hüllermeier & Waegeman, 2021). As a response to this and other problems like incompleteness of problem formalization and reward misspecification (Doshi-Velez & Kim, 2017; Klein et al., 2022; Pan et al., 2022) the field of interpretability emerged, trying to look inside the black box and understand what is actually going on. One domain of interpretability is the attempt to extract the model's uncertainty about its output. By showing such information to the user, they will be able to evaluate more easily how reliable the output is. Three metrics will be the focus of this thesis: total uncertainty, aleatoric uncertainty and epistemic uncertainty.

1.2 *Goal and Contribution*

The goal of this thesis is to examine whether Shannon entropy can be used to calculate the aleatoric and epistemic uncertainty of the output of a BERT model. By adding this subdivision of uncertainty, we not only are able to predict the model accuracy by its total uncertainty, but we also get to know why it is uncertain. This thesis evaluates through the use of metadata of a dataset whether the metrics generated through calculations derived from Shannon entropy can actually capture these uncertainties.

Our Research Questions are:

1. How can epistemic and aleatoric uncertainty be measured for hate speech classification?

2. How suitable is Shannon entropy as a measure for epistemic, aleatoric and total uncertainty for hate speech classification?

Through the answering of these research questions, this thesis aims to achieve the following contributions:

- Providing and evaluating an implementation of three uncertainty measures based on Shannon entropy for a BERT-based hate speech classifier.

- Implementing novel means to empirically evaluate whether a given metric indeed captures aleatoric or epistemic uncertainty in NLP classification.

1.3 *Structure and Method*

This thesis is structured in the following chapters: Chapter 2 includes an overview of hate speech and a primer on uncertainty calculation, ending with the research gap this thesis aims to close. Chapter 3 describes the dataset, the model architecture, its hyperparameters as well as how the evaluation methods have been implemented in both experiments. Chapter 4 describes and interprets the results of our evaluation of the uncertainty methods, and whether they match our expectations for how they should behave. Chapter 5 concludes this thesis with a discussion of the results, limitations of the approach, an outline of future work.

2 GROUNDWORK AND RELATED WORK

2.1 *Hate Speech*

What is hate speech? Although its use is ubiquitous, "[n]o formal definition exists but there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them." (Davidson et al., 2017) This leads to some scholars rejecting the 'hate' part of hate speech as misleading and non-essential, while emphasizing its multiple meanings as the cause of why there is unlikely to emerge a unified definition, instead proposing narrower definitions for each application context (Brown, 2017), such as online platforms, the legal system and academia (Sellars, 2016). In Chapter 5 we argue that dealing with this ambiguity of hate speech is critical for aleatoric uncertainty in particular. While dealing with its intricacies and dilemmas in application are beyond the scope of this thesis, we give a brief primer in Chapter 2.1.1.

The consequences of online hate speech can range from limiting the visibility of posts [2] to removal of content [3] to fines or prison sentences, if it contravenes local law. As an example, the German Criminal Code reads in Section 130(1) on Incitement of masses: [4]

> *Whoever, in a manner suited to causing a disturbance of the public peace,*
>
> *1. incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them or*
>
> *2. violates the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups, sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population*
>
> *incurs a penalty of imprisonment for a term of between three months and five years.*

2.1.1 *Hate speech detection*

If hate speech does not have a commonly accepted definition, how is detecting it even possible?

One part of the answer is that we do not need to detect *all* hate speech, but

[2] Twitter (2023), *Hateful Conduct*. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, last viewed January 2024

[3] Meta (2024), *Hate Speech*. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/, last viewed January 2024

[4] Federal Ministry of Justice (2021), *German Criminal Code*. https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html, last viewed January 2024

rather the specific kind of hate speech necessary for our context. The way we go about classifying hate speech is also going to be different, depending on whether we are talking about a legal case, disciplinary action within a company, software that automatically detects hate speech on a social media platform or some dispute in the streets. The second part of the answer is that indeed there are some things, which are often shared across all application domains. An excerpt of such a commonalities list by Sellars (2016) is:

- **Targeting of a group or individual as a member of a group**: While the protected 'groups' differ within each framework, this quality separates hate speech from other types of harmful speech.
- **Content in the message that expresses (or incites) hatred:** While some authors reject that hatred is a necessary component of hate speech (see *e.g.* (Brown, 2017)), even they agree that most types of hate speech contain some form of hostility.
- **The speech causes a harm**: Sometimes this harm strictly refers to physical harm, other times it includes more abstract modes of harm as well.
- **The speech incites bad actions beyond the speech itself**: These bad consequences can range from hatred to ethnic cleansing.
- **The context makes violent response possible**: The context of speech is crucial, especially if we are to evaluate if physical harm is a likely result of this speech. Especially on the internet, it is additionally noteworthy that speech can be transformed beyond its original context.

2.1.2 *Problems with automated hate speech detection*

Besides the general problems of creating an NLP dataset with limited resources, such as using a single data source and a limited number of qualified annotators, there are more specific difficulties in creating a hate speech dataset. Not only can hate speech be conveyed through *e.g.* sarcasm, but sometimes the broader context of a conversation is necessary for determining whether it is hate speech or not (de Gibert et al., 2018). With the reappropriation of slurs and similar through the marginalized groups (Galinsky et al., n.d.), labels can become outdated in years. What becomes labeled as hate speech or not is highly contingent on the annotator's personal background and biases, the definition given to them, and the broader context of the experiment. This is exemplified by the experiment of Ross et al. (2016), which suggests that even giving a working definition of hate speech to annotators does not erase the influence of their cultural background and personal history, but rather makes them find a middle ground between their personal feeling and the definition given to them. One proposal, therefore, is to rethink the hate speech problem as a regression problem rather than a binary yes/no classification problem (Ross et al., 2016; Sellars, 2016).

A much higher degree of caution compared to other NLP classification tasks is to be applied if existing datasets are to be used for a concrete application,

because of these factors.

Building a hate speech classifier is difficult, not just because it is difficult to fit a classifier towards the data, but also because as mentioned above, there is no universally agreed upon definition of hate speech, leading to heterogeneous datasets. Additionally, there is the danger of creating a classifier that simply labels any text mentioning minorities or using offensive words as hate speech, even if the context is benign. Google's Jigsaw has highlighted this issue by hosting an online competition with this specific theme (Jigsaw/Conversation AI, 2016). Davidson et al. (2019) did find that tweets written in African-American English (AAE) get classified as hate speech more often than tweets written in Standard American English (SAE), however they did not control for whether *e.g.* sexism may in fact be more prevalent in those tweets compared to those in SAE and actually find that *e.g.* the word "b*tch" is used three times more often in tweets written in AAE. The chosen dataset for this study addresses part of this issue by having an additional third 'offensive' class.

2.2 *Uncertainty*

Machine learning can be seen as a process of *induction*, in which a generalized model is extracted from given data points. These models cannot be provably correct, so the models themselves as well as their output are uncertain (Hüllermeier & Waegeman, 2021).

The uncertainty of a model towards an experiment can be differentiated into two kinds of uncertainties: *epistemic uncertainty* and *aleatoric uncertainty*. Epistemic uncertainty is based on our modeling error of the real life distribution and can, hence, in theory, be alleviated through better training and model architectures. Aleatoric uncertainty is based on the inherent chaotic nature of our experiments themselves such as noise and can, hence, not be alleviated through more experiment data. They are also called model uncertainty and data uncertainty (Zhou et al., 2022b) or statistical and systematic uncertainty (Hüllermeier & Waegeman, 2021) respectively. What uncertainties are regarded as reducible or irreducible are highly context-dependent: The uncertainty of a coin flip experiment can be seen as irreducible, if the coin flip is not controlled. But by building *e.g.* a coin flip robot, which always flips the coin in basically the same way, it becomes possible to reduce this uncertainty (Hüllermeier & Waegeman, 2021).

For our purposes, it can be summarized that epistemic uncertainty is mainly concerned with how out-of-distribution a data point is — "[e]pistemic Uncertainty at point x is a quantity which is high for a previously unseen x, and decreases when x is added to the training set and the model is updated" (Mukhoti et al., 2023) — while aleatoric uncertainty measures at how close a data point is to the class decision boundary (Vazhentsev et al., 2023). Aleatoric uncertainty is only meaningful in-distribution (Mukhoti et al., 2023), see also Chapter 5.2.

As to how these uncertainties are calculated, there are various methods, as indicated in Chapter 5.3.1. The author has chosen to focus on entropy-based

measures, which have become a staple in uncertainty quantification (Hüllermeier & Waegeman, 2021; Wimmer et al., 2023). The first reason for this choice is that they can be quickly added to any system, which can reasonably use dropout. The second reason being that the original paper by Smith and Gal (2018) suggested that these entropy-based measures alone could be sufficient to calculate aleatoric, epistemic, as well as total uncertainty, while the other measures focus on only one of the three.

In our use case of hate speech classification, the borderline between being rude and hate speech may be blurry for some data points — these are samples for which we expect to measure high aleatoric uncertainty, since the human annotators themselves disagree about whether to classify these data points just as offensive speech without targeting a group (characteristic) and hence not hate speech or whether they indeed do target a group (characteristic) and hence are hate speech. If a new target group emerges at test time, with *e.g.* several swear words used that have not been in the training set, we would expect to see a high epistemic uncertainty.

Uncertainty measures are also sometimes used to detect adversarial attacks (Hüllermeier & Waegeman, 2021; Koh & Liang, 2017; Smith & Gal, 2018) and can achieve state-of-the-art results in detecting out-of-distribution samples (Venkataramanan et al., 2023). While most uncertainty measures are used post hoc for already trained models, there are also approaches which propose different architectures (Hu & Khan, 2021; Huseljic et al., 2021; Sankararaman et al., 2022) or training algorithms (Tabarisaadi et al., 2022; Wei et al., 2022).

Since the total uncertainty of a model can conceptually be cleanly subdivided into epistemic uncertainty, which does decrease with additional information, and aleatoric uncertainty, which cannot be decreased with additional information, we can conveniently calculate the third metric if we have the other two, and they are scaled appropriately. As for why this may actually not work as commonly assumed, see Chapter 5.1.

2.2.1 *Shannon entropy*

A common way of calculating aleatoric (also called aleatory in *e.g.* (Kitahara et al., 2022)) and epistemic uncertainty is to calculate uncertainties metrics using Shannon entropy. Smith and Gal (2018) argue that other measures such as the estimated variance of the softmax which is used by other authors — *e.g.* in the original paper proposing using dropout for uncertainty calculation (Gal & Ghahramani, 2016) — are ad hoc and an approximation of the metrics obtained through Shannon entropy.

Shannon entropy describes the average information encoded in a probability distribution or the average amount of storage resource required, which can be approximated by the formula given in Equation 1, with K being the number of possible sequences and T the encoding sequence length.

$$\frac{logK}{T} \sim \sum_{i=1}^{n} p_i log p_i \tag{1}$$

Since the output of our model can be interpreted as "a conditional probability distribution $P(y|x)$ over some discrete set of outcomes Y" (Smith & Gal, 2018), we can calculate our **predictive entropy** $H[P(y|x)]$ as a proxy for *total uncertainty* through Equation 2:

$$H[P(y|x)] = -\sum_{y \epsilon Y} P(y|x) log P(y|x) \tag{2}$$

Our *aleatoric uncertainty* measure **expected entropy** on the other hand can be defined through Equation 3, with $\omega_i \sim q(\omega|D)$ being samples from the dropout distribution of a model whose weights and biases ω we have calculated through some dataset D:

$$E_{P(x)} H[P(Y|X)] \simeq E_{p(\omega|D)} H[p(y|x,\omega)] \simeq \frac{1}{T} \sum_{i=1}^{T} H[p(y|\omega_i, x)] \tag{3}$$

Since *epistemic uncertainty* can be calculated through simple subtraction of aleatoric uncertainty from total uncertainty by virtue of additive decomposition, we can calculate all three kinds of uncertainties from our dropout probability distributions. The resulting measure, **mutual information**, given through Equation 4 is supposed to calculate the information gain of knowing the label y of a data point x in relation to model parameters ω given a dataset D:

$$I(\omega, y|D, x) = H[P(y|x)] - E_{P(x)} H[P(Y|X)] \tag{4}$$

Since this means that we are effectively quantifying the "expected divergence of single hypotheses from the opinion given by integrating over all of them", mutual information seems in fact "not [...] well-suited to measuring a quantity taken to represent *ignorance*" (Wimmer et al., 2023), namely epistemic uncertainty.

In practice, this means we can calculate our total and aleatoric uncertainty through the code given in Code Snippet 1 and our epistemic uncertainty through simple subtraction of the latter from the former. The derivation given here is mainly based on the derivation given by Smith and Gal (2018). (For more information about the derivation and its Bayesian and second order distribution interpretation and background, see Hüllermeier and Waegeman (2021), Smith and Gal (2018), Wimmer et al. (2023), and Yu et al. (2022) and Chen (2021) for background on Shannon entropy.)

Since the calculation of Shannon Entropy requires full knowledge of the distribu-

tion of all possible models for modelling our dataset, it is usually approximated via finite-ensemble approximation. By training some number of models, *e.g.* 10, we can approximate the underlying distribution of all possible models by sampling from it (Wimmer et al., 2023). However, retraining the same model 10 times can be expensive and energy-intensive, and loading 10 different models into memory can slow down the evaluation at test time significantly. Gal and Ghahramani (2016) provide an alternative. The authors show that even this ensemble of 10 different models can be approximated by evaluating a data point 10 times by the same model with 10 different dropouts applied to the model during evaluation phase. Since by using dropout, we are sampling models from a large set of possible model configurations, they name this process Monte Carlo Dropout, also referred to as MC dropout or MCD.

Another kind of approximation would be a Bayesian Neural Network, which learns a distribution instead of deterministic weights. Each new execution of the network creates a new, somewhat randomized model, similar to running dropout during test time for non-Bayesian DL models (Mitros & Namee, 2019).

Different kinds of dropout can be used, as seen in Figure 1. One variation is whether nodes or weights are dropped, *i.e.*dropout, or dropconnect. There exist other variations, such as jumpout (Wang et al., 2019), which introduce additional normalizations, such as a consistent number of dropped values per layer of the DL network.

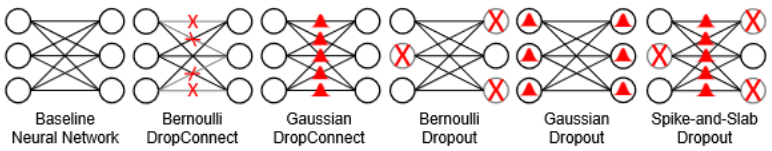

Figure 1: Different kinds of dropout (Graphic copied from (McClure & Kriegeskorte, 2022))

2.3 *Classification Architecture and Dataset choice*

In the study of D'Sa et al. (2020) a fine-tuned BERT model has vastly outperformed a CNN and Bi-LSTM, achieving an F1 score of 0.97 on binary classification of toxic vs. non-toxic speech. There is a stark drop in performance in multi-class classification, as hate speech may be difficult to differentiate from offensive speech, exacerbated by an unbalanced dataset. Since Vazhentsev et al. (2023) also report better performance of BERT as compared to ELECTRA, I have chosen BERT for our classifier.

There exists a plethora of hate speech datasets (Alatawi et al., 2020; de Gibert et al., 2018; D'Sa et al., 2020; Mathew et al., 2021; Ross et al., 2016). While Mathew et al. (2021) report an F1 score of about 0.7, which is considerably lower than the F1 score of 0.84 of D'Sa et al. (2020), this discrepancy is likely

due to the different ways the data has been selected. Not only do Mathew et al. (2021) use two data sources rather than only one, they also use a greatly expanded lexicon of hate speech keywords. This reduced performance of BERT can, therefore, be seen as an indication that epistemic uncertainty is much higher with the HateXplain dataset by Mathew et al. (2021), unless some part of the training process is making a significant impact instead. Furthermore, during the annotation process, the annotators have provided the target groups of hate speech, if applicable. In Chapter 4.2 we use this information to subdivide our dataset for our epistemic uncertainty evaluation.

Usually we are not given some ground truth uncertainty for a data point. Since in the HateXplain dataset we have the ground truth annotations of all three annotators, this enables us to make use of the disagreement between annotators in labeling as a proxy for aleatoric uncertainty of that data point. Without the ground truth annotations by each annotator, we could not evaluate aleatoric uncertainty through our method in Chapter 4.1.

The HateXplain dataset and code basis (Mathew et al., 2021) also already implements LIME with faithfulness and plausibility metrics for BERT and other models, making it easy to build on for further explainability research.

And as a last, somewhat subjective point in favor of HateXplain, I subsampled a couple of dozen data points of different datasets to inspect the annotations. The annotators at the HateXplain dataset seemed to me to have annotated in the most reliable and consistent way.

In summary, since HateXplain has been presented as a benchmark dataset for hate speech with a particular focus towards explainable machine learning, and its enabling of both of our experiments in Chapter 3 through additional information, this thesis uses HateXplain as its dataset.

2.4 *Research Gap*

In Hüllermeier and Waegeman (2021), the authors attempt to sketch the current landscape of aleatoric and epistemic uncertainty quantification research in a machine learning setting. Shannon entropy holds a special place there, since it can be justified axiomatically, even though there are also theoretical problems with it (Wimmer et al., 2023). However, the author could not find an implementation of uncertainty metrics based on Shannon entropy for NLP classification.

Various epistemic uncertainty metrics and one aleatoric uncertainty metric have been applied for hate speech classification in Vazhentsev et al. (2023). The evaluation method used for these uncertainty metrics is the same as it would be for total uncertainty, looking at the correlation between performance and uncertainty. While they demonstrate that it is possible to combine one epistemic and one aleatoric uncertainty estimation technique to obtain a hybrid uncertainty quantification metric which outperforms the state of the art of total uncertainty metrics, they only evaluate the combined total uncertainty of both

metrics and assume that their single metrics actually calculate the uncertainty that they are supposed to. In their paper they note that their "base epistemic [uncertainty evaluation] methods sometimes cannot outperform even the weak SR baseline" and that "[t]his effect might appear because the majority of model mistakes arise from ambiguity rather than [out-of-distribution] instances", since softmax response (SR) is an aleatoric uncertainty estimation method. The fact that this is just their assumption rather than being able to investigate this matter demonstrates the need and current lack of approaches to gauge the actual epistemic and aleatoric uncertainty for a data point in NLP classification.

As an example, Shannon entropy with its derived measures of expected entropy and mutual information has been presented as a way to calculate aleatoric and epistemic entropy respectively. Even if there have been suspicions about it underestimating uncertainty during its publication (Smith & Gal, 2018) and prior theoretical work suggesting that they are not the same (Dubois & Hüllermeier, 2007), it hasn't been until Wimmer et al. (2023) that we have a clear understanding of why that is. This situation of ambivalence and various investigations into the efficacy of mutual information as epistemic uncertainty could have been prevented (Bertoni et al., 2019; Mukhoti & Gal, 2019; Nair et al., 2020; Smith & Gal, 2018; Zhou et al., 2022a), if a reliable method of evaluating epistemic and aleatoric uncertainty had been available.

In Table 1, copied from (Xiao & Wang, 2019), we can *e.g.* see how by considering mutual information, their baseline performance actually decreases, similar to our results in 4.2.

Model	**Yelp 2013**	**Yelp 2014**	**Yelp 2015**	**IMDB**
(RGS MSE)				
Baseline	0.71	0.72	0.72	3.62
Baseline + MU	0.57	0.55	0.55	3.20
Baseline + DU	0.84	0.75	0.73	3.74
Baseline + both	**0.57**	**0.54**	**0.53**	**3.13**
Relative Improvement (%)	19.7	25.0	26.4	13.5

Table 1: Baseline + model uncertainty (MU) or epistemic uncertainty in the form of mutual information performs worse than the baseline

For image classification, establishing such an evaluation method is intuitive (Huseljic et al., 2021; Venkataramanan et al., 2023). An evaluation method of whether the given aleatoric and epistemic uncertainty metrics for NLP classification actually measure what they are supposed to measure has been missing.

In conclusion, our literature review could find neither a published method for evaluating aleatoric and epistemic uncertainty as such for NLP classification, nor could we find an implementation of aleatoric and epistemic uncertainty measures for an NLP classifier based on Shannon entropy. This research gap guides the research questions we seek to answer, as seen in Chapter 1.2.

3 IMPLEMENTATION

The code[5] is mainly based on the implementations of Mathew et al. (2021) for accessing the HateXplain database[6] and Smith and Gal (2018) for the formulae and implementation of expected entropy, predictive entropy and mutual information[7]. By making use of dropout during test time, we can calculate total, aleatoric and epistemic uncertainty metrics based on Shannon entropy of our fine-tuned BERT model, as seen in Code Snippet 1.

```
def entropy(probability_array):
    entropy_sum = 0
    for i in range(len(probability_array)):
        entropy_sum += -probability_array[i] * math.log(probability_array[i],2)
    return entropy_sum

# Total Uncertainty
def predictive_entropy(batch_probabilities):
    pred_entropy = []
    for data_point in batch_probabilities:
        mean_prob = np.mean(data_point, axis=1)
        pred_entropy.append(entropy(mean_prob))
    return pred_entropy


# Aleatoric Uncertainty
def expected_entropy(batch_probabilities):
    exp_entropy = []
    for data_point in batch_probabilities:
        point_entropy=0
        for dropout in range(len(data_point)):
            point_entropy += entropy(data_point[dropout])
        exp_entropy.append(point_entropy/len(data_point))
    return exp_entropy


#Epistemic Uncertainty
def mutual_info(pred_entropy=0, exp_entropy=0):
    if len(pred_entropy) != len(exp_entropy):
        print("Entropy arrays do not have the same length!")
        return 0
    return np.subtract(pred_entropy, exp_entropy)
```

Code Snippet 1: By using our formulae from Chapter 2.2.1, we can calculate our total, aleatoric and epistemic uncertainty metrics.

[5]The code will be available at https://github.com/Pyrphoros42/hateBERT-hybrid-uncertainty
[6]https://github.com/hate-alert/HateXplain
[7]https://github.com/lsgos/uncertainty-adversarial-paper

3.1 *Model architecture and dataset*

We use the 'bert-base-uncased' model as our base model and fine-tune it over three epochs. For our experiments, we use the best-performing hyperparameters as calculated by the HateXplain authors[8], with the difference being reducing the epoch number from 20 to 3. This achieves basically the same performance on the test data, with accuracy being 0.01 lower. This suggests to me that the authors of HateXplain may either have been significantly overtraining or that some copying error has occurred, since in the data file of their best BERT hyperparameters, their 'model_name' is 'birnn' rather than 'bert', as seen in Figure 2. As another case in point is the manual evaluation file[9] suggesting 5 epochs and the original BERT authors (Devlin et al., 2019) using mostly 3 epochs for fine-tuning as well. The data is split into training, validation, and evaluation data according to the same 8:1:1 split as in the original HateXplain paper (Mathew et al., 2021).

3.2 *Uncertainty evaluations*

In order for us to evaluate our uncertainty metrics, we check whether the samples for which we get higher uncertainty are more likely to be wrong than the samples of lower uncertainty. Furthermore, we measure additional uncertainty metrics depending on the type of experiment. We evaluate four times with a 0.1 probability of dropout unless stated otherwise.

3.3 *Experiments*

We conduct two experiments: Experiment Dissent and Experiment Target Groups

3.3.1 *Experiment Dissent*

In **Experiment Dissent**, we are evaluating the performance of our metrics on a regularly trained model, but we are additionally checking the correlation between aleatoric uncertainty and the number of dissenting annotators. This is due to one way to think about aleatoric uncertainty being that even if annotators have full information about the sample, there can be disagreement between reasonable and informed annotators about what label to assign. The higher the aleatoric uncertainty of the sample, the higher the likelihood of disagreement, until it approaches random choice. Since the HateXplain dataset includes the way every annotator has labeled each data point, this gives us the option to look for data points, where one annotator has labeled the data point a different way from the other two or even to look at the roughly 5% of data points, where all three annotators have selected a different one of the three classes [normal, offensive, hateful]. These data points are more likely than those without dis-

[8] https://github.com/hate-alert/HateXplain/blob/master/best_model_json/bestModel_bert_base_uncased_Attn_train_TRUE.json

[9] https://github.com/hate-alert/HateXplain/blob/master/manual_training_inference.py

Figure 2: Selection from the hyperparameter file "best-Model_bert_base_uncased_Attn_train_TRUE.json"

agreement to have higher aleatoric uncertainty, which the author expected to show in an increased measure of aleatoric uncertainty with increasing number of dissenting annotators.

Since in our dataset we have given the particular annotations by each individual annotator, we can easily calculate the number of dissenting opinions, as seen in Code Snippet 2:

```
# add number of dissenting annotators to dissent_list
if(label1 != label2) && (label1 != label3) && (label2 != label3):
    annotation = label1
    count_confused+=1
    dissent_list.append(2)
elif label1 != label2 or label2 != label3:
```

```
    dissent_list.append(1)
else:
    dissent_list.append(0)
```

Code Snippet 2: Marking how big the disagreement is among annotators. (Code Snippet slightly altered)

3.3.2 *Experiment Target Groups*

Epistemic uncertainty is difficult to measure, since we need to measure how out-of-distribution some samples are compared to others in relation to our training dataset — the training dataset and test dataset are usually divided randomly from the same data source and do not have further annotations. In the HateXplain dataset, however, we have additionally target group annotations, like *African* or *Jewish*.

In **Experiment Target Groups**, we remove two target groups from the training data, through changing dataset generation process from the raw data, as seen in *e.g.* Code Snippet 3 and 4. We then evaluate at test phase whether our epistemic uncertainty measure can detect test data points targeting these groups as 'out-of-distribution'. Since our training data set ends up being about 20% smaller, we increase our training epochs from 3 to 4.

```
# take all target groups annotated at least twice
target1 = np.intersect1d(row['target1'],row['target2'])
target2 = np.intersect1d(row['target1'],row['target3'])
target3 = np.intersect1d(row['target2'],row['target3'])
target = np.union1d(np.union1d(target1,target2),target3)
# if target contains 'Hispanic' or 'Refugee' = 1, else 0
if ('Hispanic' in target) or ('Refugee' in target):
    target = 1
else:
    target = 0

```

Code Snippet 3: Marking the data point, if at least two annotators chose 'Hispanic' or 'Refugee' as a target group for it.

```
# remove target group from dataset
if params['EU']:
    X\_train = X\_train.drop(X\_train[X\_train['Target'] > 0].index)
```

Code Snippet 4: Dropping the data points from the training set, which have been annotated with 'Hispanic' or 'Refugee'.

My expectation going into this experiment was that there would be a clear difference in the epistemic uncertainty of removed and not-removed target groups during test phase.

4 EVALUATION

For total and aleatoric uncertainty we expected that with rising uncertainty, we find decreasing accuracy. In Figure 3 and 4 we can see how with increasing aleatoric and total uncertainty there is a linear decrease in average accuracy until it reaches about random chance for our three classes, *i.e.*33%.

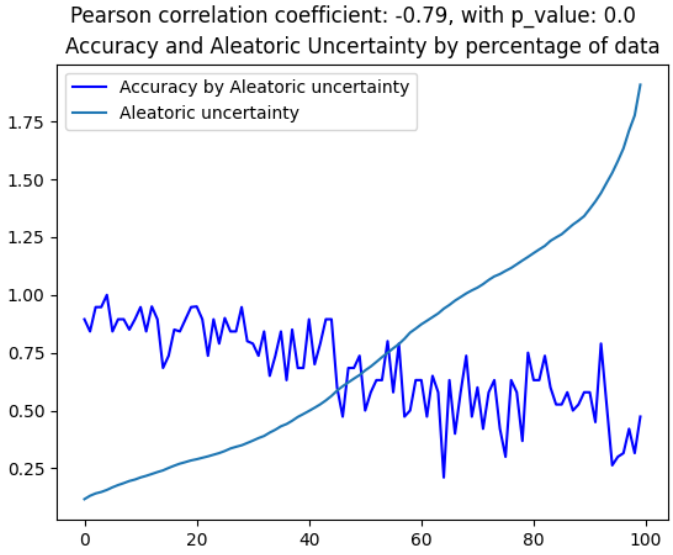

Figure 3: Correlating accuracy and aleatoric uncertainty calculation

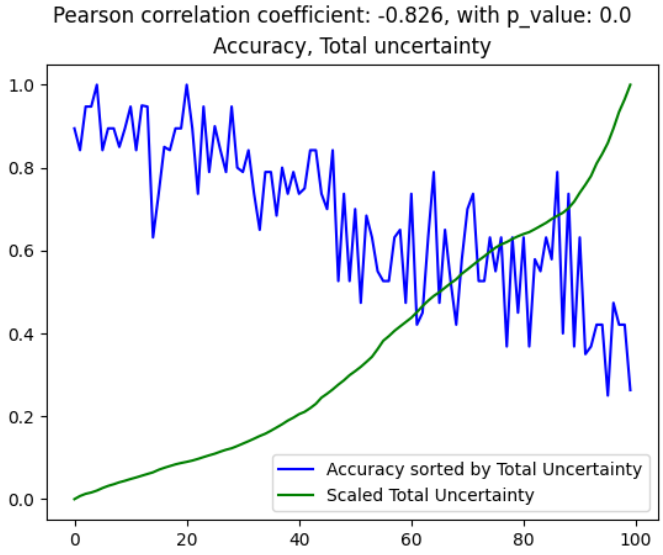

Figure 4: Correlating accuracy and total uncertainty calculation

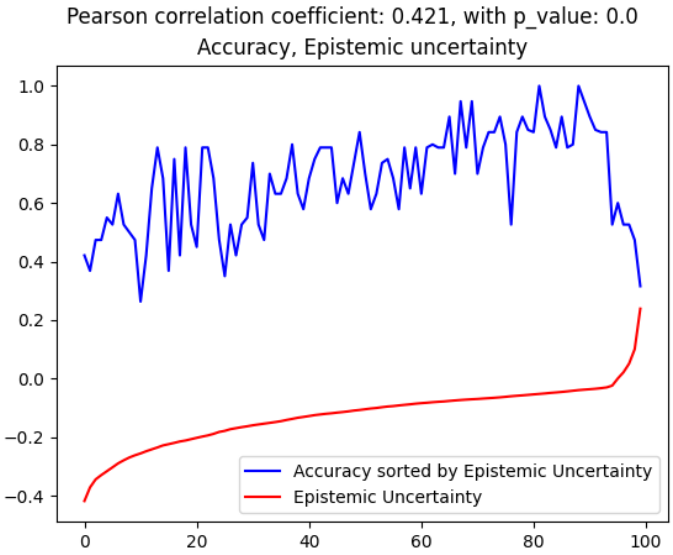

Figure 5: Correlating accuracy and epistemic uncertainty calculation

Epistemic uncertainty is a curious case. With increasing epistemic uncertainty, the accuracy of the model actually increased, as seen in Figure 5. This is likely due to the subtraction way our epistemic uncertainty is calculated: the minuend, our total uncertainty, grows slower in value size than the subtrahend, aleatoric uncertainty, so it follows that our epistemic uncertainty is positively correlated with it. Nonetheless, we see a remarkable uptick in the rise of epistemic value for the 8% of our data points with the highest epistemic uncertainty, correlating exactly with a sharp drop in accuracy, with the 1% of data points with the highest epistemic uncertainty showing an accuracy that is about the same as random choice. So while epistemic uncertainty is likely heavily underestimated, it arguably still proves a useful metric for a subset of data points.

These results stayed constant in all experiments.

Concerning higher dropout numbers, we find that increasing the percentage of dropout can actually significantly deteriorate the baseline performance of the model, as seen in Table 2. Even uncertainty metrics generally decrease in correlation, except for the case detailed in Chapter 5.1.

4.1 *Experiment Dissent*

In Experiment Dissent we took a more careful look at whether expected entropy actually captures aleatoric uncertainty.

In Figure 6 we can see in the upper plot that with increasing number of dissenting annotators our accuracy is dropping significantly by about 25 percentage points, with our accuracy at 2 dissenting annotators being 0.33, since if all three

Number of dropouts	Dropout size	Accuracy	F1 score	Precision	Recall
0	-	0.69	0.68	0.69	0.69
4	0.1	0.69	0.68	0.69	0.69
4	0.3	0.67	0.66	0.68	0.65
4	0.4	0.61	0.58	0.64	0.58
7	0.1	0.68	0.67	0.67	0.67

Table 2: A table of performance-related metrics of a trained BERT model with different dropout sizes selected

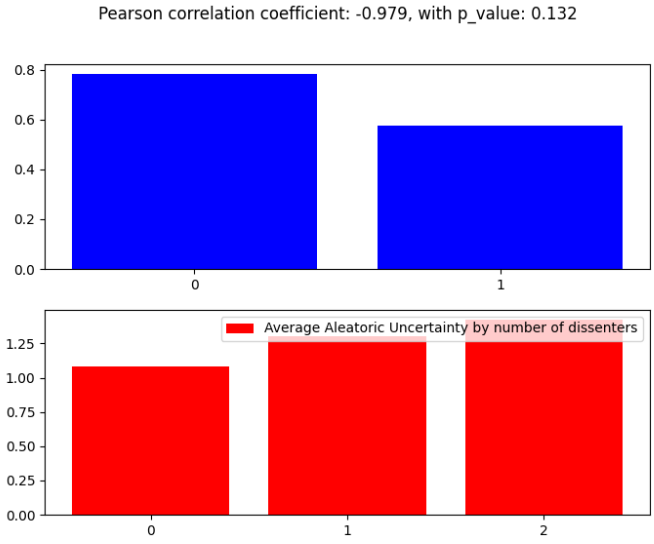

Figure 6: Upper Plot: The decrease in accuracy if there is one dissenting annotator. Lower Plot: The average amount of aleatoric uncertainty by dissenting annotators.

annotators disagree, no class can be reasonably selected. In the lower plot we can see that as expected, our average aleatoric uncertainty rises with the number of dissenting annotators, giving us a Pearson correlation size of -0.979. As the number of single data points is low, the p_value is relatively high still, compared to the extremely high correlation score. It is worth pointing out however that total uncertainty shows the same kind of strong correlation to the number of dissenters as aleatoric uncertainty, as seen in Appendix Figure 9.

4.2 *Experiment Target Groups*

We do not find the expected connection between higher mutual information and a hate speech data point with an unknown target group. While accuracy is significantly lower in bars one and three in Figure 7, epistemic uncertainty stays roughly the same.

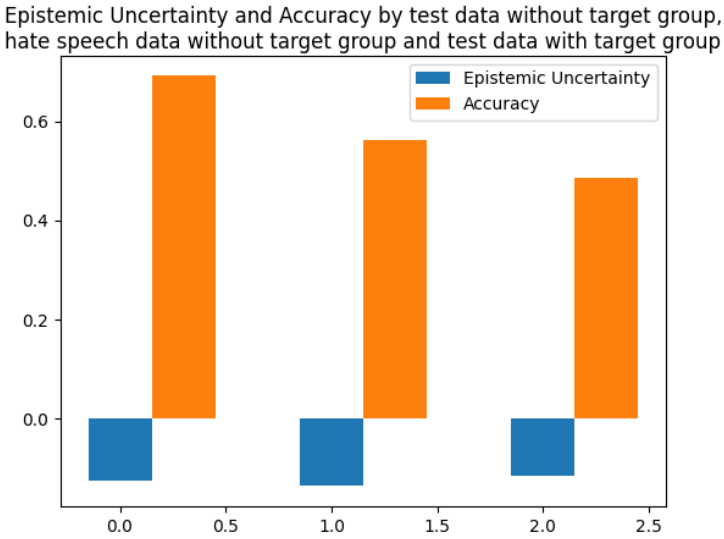

Figure 7: No relationship found between epistemic uncertainty and a data point targeting an unknown target group

Repeating the experiment with an increased number of dropouts from 4 to 7 did not change this relationship, suggesting that our calculated mutual information metric does not in fact capture epistemic uncertainty. In order to make sure that epistemic uncertainty wasn't somehow encoded in both expected entropy and predictive entropy and canceled out through the subtraction, we evaluated both metrics for epistemic uncertainty, but did not find that they increased for data points with unknown target groups, see Appendix Figures 13 and 14.

For a more detailed discussion of these and other results of this experiment, see the following Chapter 4.3.

4.3 *Interpretation of results*

While the author could not find a case of uncertainty metrics using Shannon entropy being used in BERT for NLP, it has been surprising to them how big the correlation between the aleatoric/total uncertainty measures and accuracy turned out to be. The effect is pronounced enough that it could be reasonably applied during deployment, by *e.g.* requiring no human oversight for classification below an uncertainty threshold value and rejecting classification above a different, higher threshold value (see also: Herbei and Wegkamp (2006)).

As it has been pointed out in 4.1, it is not just aleatoric uncertainty showing a high correlation with the number of dissenting annotators, but also total uncertainty. Since our total uncertainty calculation barely captures epistemic uncertainty and instead mostly consists of aleatoric uncertainty, this is not a surprising result.

Our approach of ranking data points by metrics is the same as applied in (Vazhentsev et al., 2023), where it is a crucial step in normalizing different metrics to later extract a hybrid metric. Similarly, our mode of analysis of uncertainty metrics does not depend on the values that the metrics end up taking, if they were to be replaced by other metrics.

It remains unclear, whether with smarter ways of removing certain types of data from the training set it might still be possible to find a correlation between mutual information and accuracy for out-of-distribution samples. Although "mutual information behaves as expected but does not seem well suited to measuring a quantity taken to represent *ignorance*" (Wimmer et al., 2023), it may still be possible to find a small correlation with a more suitable experiment.

Since in this study the same dataset has been separated into training and test data, it is reasonable to assume that the out-of-distribution-ness for new kinds of test data will be higher than our test data. So while our techniques have been very performant, it is to be expected that techniques focused on epistemic uncertainty as in (Mukhoti et al., 2023; Vazhentsev et al., 2023) will relatively rise in importance, making up 50% of the final total uncertainty calculation of their "hybrid uncertainty quantification". This is especially important since mutual information has been shown in various previous works (Smith & Gal, 2018; Wimmer et al., 2023) and this thesis to not be a good measure of epistemic uncertainty: mutual information encodes the divergence of a single hypothesis compared to the integration over all of them, rather than the information gain. Relatedly, predictive entropy, our measure of total uncertainty decreasing does not necessarily mean that informedness is increasing. In fact, predictive entropy stays the same for data points with unknown target groups, just as in the case of mutual information, as seen in Appendix Figure 14.

5 DISCUSSION, LIMITATIONS, FUTURE WORK

5.1 *Discussion and Limitations*

In general, our model has the same limitations as the model of Mathew et al. (2021): Lacking context like post history of users and the conversation and the exclusive focus on content in the English language. Evaluations by Ovadia et al. (2019) suggest that uncertainty metrics which perform well on test sets may break down on data with a significant shift, so this remains to be tested as well. This technique is post hoc and can be easily combined with other explainable machine learning techniques. Increasing dropout size quickly leads to deteriorating classification metrics (see Figure 2).

One factor that has been considered by ECRI (2016), is the danger of discourse within a minority group to be classified as hate speech, *e.g.* if an African American calls another one the n-word, without intending a racist disparagement. It remains to be examined whether this discourse is at *disproportionate* risk to be falsely labeled as hate speech. Disturbingly, Davidson et al. (2019) find increased classification of African-American English compared to Standard American English as hate speech, although it remains to be seen, whether this is due to increased prevalence of hate speech or some other factor.

Another critique of entropy-based calculation of uncertainty comes from a group of authors including Yarin Gal (Mukhoti et al., 2023), one of the two original authors of the paper introducing this measure in Smith and Gal (2018), concluding that "[i]n practice, however, these methods are either unable to scale to large datasets and model architectures, suffer from low uncertainty quality, or require expensive Monte-Carlo sampling". Although our technique does not require any special training or architecture, the evaluation of test data does take significantly more time. It is, therefore, worth pointing out, that other ways of using MC Dropout to simulate entropy exist, which only change the output layer, reducing the overhead to less than 0.5%, while maintaining mediocre performance according to (Vazhentsev et al., 2022), getting rid of one of the significant drawbacks of this technique.

Perfect additive decomposition, which has been assumed in most of the research, may not hold true, argue Wimmer et al. (2023). While an ideal epistemic uncertainty measure could calculate the epistemic uncertainty at any given informedness of the model, a faithful aleatoric uncertainty measure is reliant on perfect information about the decision boundaries, as illustrated in Figure 8.

Since this opens up another possible interpretation as to why aleatoric uncertainty is dominant in total uncertainty, *i.e.*that there is already a high degree of informedness, it is crucial to evaluate epistemic uncertainty metrics on data of which we know it has a low degree of informedness about it as well as on data of which we know it has a high degree of informedness about it. When calculating aleatoric uncertainty from total uncertainty and epistemic uncertainty (as many techniques in Chapter 5.3.1 do), care should be given that this estimation is likely

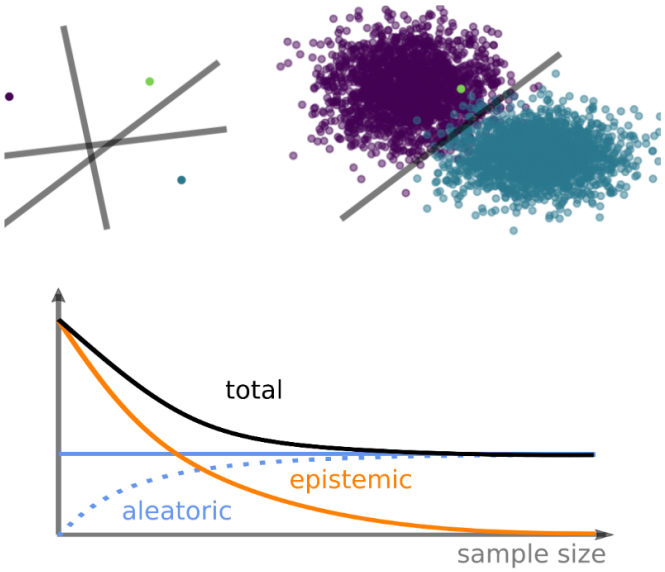

Figure 8: Epistemic uncertainty decreases over the training process and approaches 0, while as the aleatoric uncertainty metric starts low and approaches the true aleatoric uncertainty boundary. (Graphic copied from Wimmer et al. (2023))

to be a lower bound on the true aleatoric uncertainty, especially if informedness is low.

Overall, "the field [of uncertainty quantification] develops very dynamically and is far from being settled [with n]ew proposals for modeling and quantifying uncertainty appear[ing] on a regular basis [...]" (Hüllermeier & Waegeman, 2021)

5.2 *Uncertainty and explainable machine learning*

How can explainable machine learning methods be evaluated? Here Doshi-Velez and Kim (2017) and Madsen et al. (2022) provide three different approaches:

1. Application-grounded: Does the explanation improve the performance of the Human-Computer system in real world application?

2. Human-grounded: Are the explanations useful to humans? Do they *e.g.* help them to predict the output of the model or perform outlier detection?

3. Functionality-grounded: Also known as faithfulness or fidelity, how well does the explanation represent the underlying model mechanisms?

Within this framework, our evaluation of uncertainty as well as uncertainty itself would be purely functionality-grounded. This seems reasonable for our evaluation of uncertainty, where we are most concerned about whether the metric actually evaluates what it is supposed to evaluate, but its real-world benefits remain unproven.

While the purpose of measuring uncertainty is to gain additional information about the output of the model, we purposefully avoid calling our uncertainty metrics 'explanations' in this thesis, since they are not *explaining* model behavior or causes for classification. Instead, we are measuring *e.g.* to what degree certain behavior of the model has been determined randomly or through training in the case of epistemic uncertainty. These measurements cannot on their own provide insight into what factors have led to this model classifying the data point in this particular way, or what factors of the data point would need to change in order for it to be classified differently. Rather than explaining the causal connection between the data point, model and output, total uncertainty is measuring the degree of *lack* of causal connection between the data point, model, and output. With subdividing total uncertainty into aleatoric and epistemic uncertainty, we gain the ability to look at the degree of lack of causal connection between data point and output dependent on the relation of the data point to the class division space of the model (aleatoric uncertainty) or the informedness about the region of state space of the model in the data point region (epistemic uncertainty).

Uncertainty is often not considered in explainable machine learning surveys (see *e.g.* Adadi and Berrada (2018) and Molnar (2020). Besides the reasons given above, one reason for this may partly be due to it is more so being seen as a method of a different domain, like out-of-distribution detection (Venkataramanan et al., 2023) or performance enhancing optimization (Herbei & Wegkamp, 2006). It can also be seen as a reliability estimation score tailored to specific instances, making it arguably more powerful than aggregated reliability scores in specific use cases (Hüllermeier & Waegeman, 2021). This enables the system to reject individual samples when predicted reliability is low, such as in Herbei and Wegkamp (2006). A further reason is the conception that "humans struggle to deal with uncertainty" (Doshi-Velez & Kim, 2017; Došilović et al., 2018) and, therefore, arguably not being a good explanation of the model to humans, even if it is functionality-grounded. This assumption has not been tested for, to the knowledge of the author. As the research field of uncertainty will continue to develop, it may be the case that new, non-probabilistic conceptions of uncertainty will emerge which will make uncertainty a lot more understandable by humans (Dubois et al., 1996; Helton et al., 2004).

Lastly, it is worth mentioning that there is some research emerging trying to bridge the gap between uncertainty and explainability research (Thuy & Benoit, 2023), but this research remains functionality-grounded thus far and, therefore, is not addressing the issue about whether and if so how this measure will actually bring utility to its users.

5.3 *Research Questions*

5.3.1 *How can epistemic and aleatoric uncertainty be measured for hate speech classification?*

Shannon entropy is one of the most commonly used methods for determining total, aleatoric and epistemic uncertainty (Hüllermeier & Waegeman, 2021; Wimmer et al., 2023) and is detailed in Chapter 2.2.1 and Chapter 3. Our final but others exist:

- **Epistemic uncertainty:**
  - Mukhoti et al. (2023) present a novel approach using features extracted from a Gaussian Mixture Model, trained with the predictions of the original model, to calculate epistemic uncertainty specifically. They call this approach *Deep Deterministic Uncertainty*.
  - Vazhentsev et al. (2023) apply several uncertainty quantification techniques which previously were only used on image recognition on hate speech classification: *Mahalanobis Distance* and *Robust Density Estimation*.
  - Sale et al. (2023) use *Credal Sets* to determine epistemic uncertainty. They present strong theoretical arguments as to why epistemic uncertainty calculated through Credal Sets are a sensible measure for binary, but not multi-class use cases, so it is only of limited use for most hate speech classifiers, which have three class labels.
- **Aleatoric uncertainty:**
  - Vazhentsev et al. (2023) use *Softmax Response* for calculating aleatoric uncertainty, which similar to Shannon entropy makes use of the output of the output layer.
  - Kull et al. (2017) improve on previous *Calibration* metrics, which do not differentiate between epistemic and aleatoric uncertainty, but mostly measure aleatoric uncertainty (Hüllermeier & Waegeman, 2021). Their experiment uses a different, difficult NLP classification task.
  - In the experiments of Valdenegro-Toro and Mori (2022), *Ensembles* produce slightly better aleatoric uncertainty performances than using dropout with a single model. The obvious cost is having to train multiple models and evaluating them, instead of only training one model. While their experiments have not been done on hate speech, it is likely that this connection will still hold, as "ensembling can be seen conceptually as a way of sampling", with this process being more *IID* than using dropout.

Additionally, epistemic and aleatoric uncertainty can also be calculated through

non-probabilistic uncertainty conceptions such as evidence theory, possibility theory or interval analysis (Helton et al., 2004) or be applied to architectures different from DL models, such as random forests (Shaker & Hüllermeier, 2020).

In this thesis, we have not just implemented new uncertainty quantification metrics for NLP classification based on Shannon entropy, but more crucially developed approaches to evaluate whether our aleatoric and epistemic uncertainty metrics actually measure the type of uncertainty they are supposed to measure.

5.3.2 *How suitable is Shannon entropy as a measure for epistemic, aleatoric and total uncertainty for hate speech classification?*

As seen in Chapter 4 and Chapter 5.1, using predictive and expected entropy as total and aleatoric uncertainty metrics yields measurements that not only show a strong correlation with overall accuracy, but the these metrics also increase with uncertainty in the data generation process, as they should.

We could not validate mutual information or predictive entropy as an epistemic uncertainty measure, mirroring recent work like Vazhentsev et al. (2023). In fact, our evaluation suggests that mutual information barely behaves in the way we expect it to, with the possible exception of the roughly 5% highest uncertainty samples. We can, therefore, assume that our predictive entropy measure for total uncertainty generally underestimates epistemic uncertainty as well, and this is in fact what we find in Appendix Figure 14.

It is also worth mentioning that using dropout over the whole model increases evaluation time by the number of times dropout is used. If evaluation speed is crucial during deployment, it may be more suitable to use a variation which only uses dropout on the last layer, while suffering some performance loss, like *e.g.* the "diverse determinantal point process" variation by Vazhentsev et al. (2022).

5.4 *Outlook and Future Work*

A straight-forward way to try to improve the performance of our system is to apply regularization, like *e.g.* spectral normalization (Vazhentsev et al., 2023). This should be done with caution, however, as regularizations, have shown to sometimes significantly improve or worsen performance of uncertainty metrics of NLP classifiers in (Vazhentsev et al., 2022).

The author's approach of using disagreement between annotators as a proxy for aleatoric uncertainty compensates for the lack of ground truth uncertainty usually present in data sets. To the best of the author's knowledge, this way of using metadata in the dataset has not been done before and can serve as a practical way to evaluate aleatoric uncertainty performance on any NLP classification task, no longer needing to rely on testing new aleatoric uncertainty techniques with image classification, whose uncertainty evaluation is more

obvious, but potentially not directly transferable (Hüllermeier & Waegeman, 2021).

One way in which dataset creators could make the evaluation of aleatoric uncertainty measures more reliable is by having more annotators per sample. Since annotation is expensive, these additional annotators could be *e.g.* only assigned to those samples, where there is already at least one dissenting annotator, saving some resources by focusing them on those samples, that are of particular interest to the evaluation of aleatoric uncertainty and general quality control.

Epistemic uncertainty supports active learning, by providing a metric with data point during deployment, which as indicated by Nguyen et al. (2021), can show how much information there is to learn from it. This could be useful if *e.g.* for some reason training resources are constrained, or if data points that potentially could be more effective in training are treated differently in the training process. Since epistemic uncertainty is heavily underestimated with Shannon entropy, the author assumes that hybrid uncertainty estimation work like Vazhentsev et al. (2023) which combines two different uncertainty measurement techniques — one focusing on aleatoric uncertainty and one on epistemic uncertainty — is going to prove fruitful, even without improvements to current techniques.

As with all other methods, this approach gets more effective the higher the quality of the data. However, it is more crucial that *all* annotators are working reliably in order for our dissent metric to be useful, whereas for most other tasks it is sufficient that the majority of annotators are working reliably. In view of this, it is significant that Ross et al. (2016) show that there is high disagreement between annotators as to what constitutes hate speech and how to classify particular samples. Even after providing expert annotators a definition, which they should apply, they still had a low degree of agreement. The authors argue that this demonstrates not only the significance of personal background and attitudes of the annotators — making their selection process more crucial, but also the emphasizing need for "a new coding scheme which includes clear-cut criteria that let people distinguish hate speech from other content", which they have been working since 2016, but published anything related since, highlighting the difficulties that such an endeavor entails, many of which are not technical in nature. Calculating aleatoric uncertainty offers a technical step forward towards another one of their proposed solutions to this problem: calculating a "degree of hatefulness" rather than treating hate speech classification as a binary yes-or-no classification task may not only reflect the sentiments of human annotators better, but also make the machine learning system more portable, as different platforms may have different levels of acceptability of hatefulness. Sellars (2016) explicitly encourages a confidence-based measure of hate speech as well. Ironically it is Facebook that has implemented a tiered hate speech system,[10] so even in this case, where there have been a lot of scandals, research and public pressure does seem to pay off. By making better tools and concepts widespread, it is more difficult for reluctant stakeholders to not make

[10]Meta (2024), *Hate Speech*. https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/, last viewed January 2024

an effort.

As a last point, the author would like to remind the reader that technical solutions are never sufficient for tackling broad societal problems like hate speech and should always be applied in conjunction with other approaches like education programs for various groups in society, promoting awareness-raising and counter-speech, civil and criminal liability, providing psychological and legal assistance for victims, etc. (ECRI, 2016) and be conceptualized with related trends like echo chambers (Cinelli et al., 2021), online radicalization (Koehler, 2014), complex changes to identity due to globalization (Bornman, 2003) and loss of trust in mass media, democratic institutions and the economic future (Hosking, 2019).

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI) [Conference Name: IEEE Access]. *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2020, October 1). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. https://doi.org/10.48550/arXiv.2010.00357

Bertoni, L., Kreiss, S., & Alahi, A. (2019). Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Bornman, E. (2003). Struggles of identity in the age of globalisation : Theory : Research article [Publisher: UNISA Press]. *Communicatio : South African Journal of Communication Theory and Research*, *29*(1), 24–47. https://doi.org/10.10520/EJC27821

Brand, M. (2020, October 27). *Facebook is tilting the political playing field more than ever, and it's no accident* [The conversation]. Retrieved June 1, 2023, from http://theconversation.com/facebook-is-tilting-the-political-playing-field-more-than-ever-and-its-no-accident-148314

Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468. https://doi.org/10.1007/s10982-017-9297-1

Chen, R. X. F. (2021, April 23). A brief introduction to shannon's information theory. https://doi.org/10.48550/arXiv.1612.09316

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., … Sculley, D. (2020, November 24). Underspecification presents challenges for credibility in modern machine learning. https://doi.org/10.48550/arXiv.2011.03395

Davidson, T., Bhattacharya, D., & Weber, I. (2019, May 29). Racial bias in hate speech and abusive language detection datasets. https://doi.org/10.48550/arXiv.1905.12516

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, March 11). Automated hate speech detection and the problem of offensive language. https://doi.org/10.48550/arXiv.1703.04009

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018, October). Hate speech dataset from a white supremacy forum. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd workshop on abusive language online (ALW2)* (pp. 11–

20). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5102

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Doshi-Velez, F., & Kim, B. (2017, March 2). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

D'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, 1–5. https://doi.org/10.1109/OCTA49274.2020.9151853

Dubois, D., Prade, H., & Smets, P. (1996). Representing partial ignorance [Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans]. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *26*(3), 361–377. https://doi.org/10.1109/3468.487961

Dubois, D., & Hüllermeier, E. (2007). Comparing probability measures using possibility theory: A notion of relative peakedness. *Int. J. Approx. Reasoning*, *45*(2), 364–385. https://doi.org/10.1016/j.ijar.2006.06.017

ECRI. (2016). *ECRI general policy recommendation n°15 - european commission against racism and intolerance (ECRI) - www.coe.int* [European commission against racism and intolerance (ECRI)]. Retrieved January 21, 2024, from https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

Feiner, J., & Bursztynsky, L. (2021, September 14). *Facebook documents show how toxic instagram is for teens, wall street journal reports* [CNBC]. Retrieved June 1, 2023, from https://www.cnbc.com/2021/09/14/facebook-documents-show-how-toxic-instagram-is-for-teens-wsj.html

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*. Retrieved March 20, 2023, from https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text

Galinsky, A. D., Hugenberg, K., Groom, C., & Bodenhausen, G. V. (n.d.). The reappropriation of stigmatizing labels: Implications for social identity. In *Identity issues in groups* (pp. 221–256). Emerald (MCB UP). https://doi.org/10.1016/s1534-0856(02)05009-0

Helton, J. C., Johnson, J. D., & Oberkampf, W. L. (2004). An exploration of alternative approaches to the representation of uncertainty in model

predictions. *Reliability Engineering & System Safety*, *85*(1), 39–71. https://doi.org/10.1016/j.ress.2004.03.025

Herbei, R., & Wegkamp, M. H. (2006). Classification with reject option [Publisher: [Statistical Society of Canada, Wiley]]. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *34*(4), 709–721. Retrieved August 23, 2023, from https://www.jstor.org/stable/20445230

Hosking, G. (2019, July 11). The decline of trust in government [Section: Trust in Contemporary Society]. In *Trust in contemporary society* (pp. 77–103). Brill. https://doi.org/10.1163/9789004390430_007

Hu, Y., & Khan, L. (2021). Uncertainty-aware reliable text classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 628–636. https://doi.org/10.1145/3447548.3467382

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Huseljic, D., Sick, B., Herde, M., & Kottke, D. (2021). Separation of aleatoric and epistemic uncertainty in deterministic deep neural networks [ISSN: 1051-4651]. *2020 25th International Conference on Pattern Recognition (ICPR)*, 9172–9179. https://doi.org/10.1109/ICPR48806.2021.9412616

Jigsaw/Conversation AI. (2016). *Jigsaw unintended bias in toxicity classification | kaggle*. Retrieved December 17, 2023, from https://web.archive.org/web/20231217130814/https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Kitahara, M., Bi, S., Broggi, M., & Beer, M. (2022). Nonparametric bayesian stochastic model updating with hybrid uncertainties. *Mechanical Systems and Signal Processing*, *163*, 108195. https://doi.org/10.1016/j.ymssp.2021.108195

Klein, L., El-Assady, M., & Jäger, P. F. (2022, July 11). From correlation to causation: Formalizing interpretable machine learning as a statistical process. https://doi.org/10.48550/arXiv.2207.04969

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the internet [Number: 1]. *Journal for Deradicalization*, (1), 116–134. Retrieved January 21, 2024, from https://journals.sfu.ca/jd/index.php/jd/article/view/8

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions [ISSN: 2640-3498]. *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894. Retrieved July 14, 2023, from https://proceedings.mlr.press/v70/koh17a.html

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity*. Retrieved July 28, 2023, from https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kull, M., Filho, T. S., & Flach, P. (2017, 20–22 Apr). Beta calibration: a well-founded and easily implemented improvement on logistic calibration

for binary classifiers. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th international conference on artificial intelligence and statistics* (pp. 623–631, Vol. 54). PMLR. https://proceedings.mlr.press/v54/kull17a.html

Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, *55*(8), 155:1–155:42. https://doi.org/10.1145/3546577

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection [Number: 17]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

McClure, P., & Kriegeskorte, N. (2022). Representing inferential uncertainty in deep neural networks through sampling. Retrieved April 3, 2023, from https://openreview.net/forum?id=HJ1JBJ5gl

Menn, J. (2015). EXCLUSIVE: Ex-employees: Russian antivirus firm faked malware to harm rivals. *Reuters*. Retrieved April 3, 2023, from https://www.reuters.com/article/kaspersky-rivals-idINKCN0QJ1D520150814

Merrill, J. B., & Oremus, W. (2021, October 26). *Five points for anger, one for a 'like': How facebook's formula fostered rage and misinformation* [Washington post] [Section: Technology]. Retrieved June 1, 2023, from https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Mitros, J., & Namee, B. M. (2019). On the validity of bayesian neural networks for uncertainty estimation.

Molnar, C. (2020). *Interpretable machine learning* [Google-Books-ID: jBm3DwAAQBAJ]. Lulu.com.

Moore, G. E. (1965). *Cramming more components onto integrated circuits*. Retrieved April 2, 2023, from https://web.archive.org/web/20190327213847/https://newsroom.intel.com/wp-content/uploads/sites/11/2018/05/moores-law-electronics.pdf

Mukhoti, J., & Gal, Y. (2019). Evaluating bayesian deep learning methods for semantic segmentation.

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H. S., & Gal, Y. (2023). Deep deterministic uncertainty: A new simple baseline, 24384–24394. Retrieved January 8, 2024, from https://openaccess.thecvf.com/content/CVPR2023/html/Mukhoti_Deep_Deterministic_Uncertainty_A_New_Simple_Baseline_CVPR_2023_paper.html

Nadeem, R. (2020, February 21). *3. concerns about democracy in the digital age* [Pew research center: Internet, science & tech]. Retrieved June 1, 2023, from https://www.pewresearch.org/internet/2020/02/21/concerns-about-democracy-in-the-digital-age/

Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, *59*, 101557. https://doi.org/https://doi.org/10.1016/j.media.2019.101557

Nguyen, V.-L., Shaker, M. H., & Hullermeier, E. (2021, June). How to measure uncertainty in uncertainty sampling for active learning - machine learning. https://link.springer.com/article/10.1007/s10994-021-06003-9#citeas

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model' s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, *32*. Retrieved January 25, 2024, from https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

Pan, A., Bhatia, K., & Steinhardt, J. (2022, February 14). The effects of reward misspecification: Mapping and mitigating misaligned models. https://doi.org/10.48550/arXiv.2201.03544

Quinn, T. (2013, March 25). *Introducing hatebase: The world's largest online database of hate speech*. Retrieved November 17, 2023, from https://thesentinelproject.org/2013/03/25/introducing-hatebase-the-worlds-largest-online-database-of-hate-speech/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training [Publisher: OpenAI].

Rosenberg, S. P. (2012). Genocide is a process, not an event [Publisher: University of Toronto Press]. *Genocide Studies and Prevention*, *7*(1), 16–23. https://doi.org/10.3138/gsp.7.1.16

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, 6–9. Retrieved November 17, 2023, from https://www.research.ed.ac.uk/en/publications/measuring-the-reliability-of-hate-speech-annotations-the-case-of-

Sale, Y., Caprio, M., & Höllermeier, E. (2023). Is the volume of a credal set a good measure for epistemic uncertainty? [ISSN: 2640-3498]. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 1795–1804. Retrieved January 26, 2024, from https://proceedings.mlr.press/v216/sale23a.html

Sankararaman, K. A., Wang, S., & Fang, H. (2022, June 1). BayesFormer: Transformer with uncertainty estimation. https://doi.org/10.48550/arXiv.2206.00826

Sellars, A. (2016, December 1). Defining hate speech. https://doi.org/10.2139/ssrn.2882244

Shaker, M. H., & Hüllermeier, E. (2020). Aleatoric and epistemic uncertainty with random forests. In M. R. Berthold, A. Feelders, & G. Krempl (Eds.), *Advances in intelligent data analysis xviii* (pp. 444–456). Springer International Publishing.

Smith, L., & Gal, Y. (2018, March 22). Understanding measures of uncertainty for adversarial example detection. https://doi.org/10.48550/arXiv.1803.08533

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks [Conference Name: IEEE Transactions on Evolutionary

Computation]. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

Tabarisaadi, P., Khosravi, A., Nahavandi, S., Shafie-Khah, M., & Catalão, J. P. S. (2022). An optimized uncertainty-aware training framework for neural networks [Conference Name: IEEE Transactions on Neural Networks and Learning Systems]. *IEEE Transactions on Neural Networks and Learning Systems*, 1–8. https://doi.org/10.1109/TNNLS.2022.3213315

Thuy, A., & Benoit, D. F. (2023). Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2023.09.009

Valdenegro-Toro, M., & Mori, D. S. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1508–1516. https://doi.org/10.1109/CVPRW56347.2022.00157

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. Retrieved April 3, 2023, from https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vazhentsev, A., Kuzmin, G., Shelmanov, A., Tsvigun, A., Tsymbalov, E., Fedyanin, K., Panov, M., Panchenko, A., Gusev, G., Burtsev, M., Avetisian, M., & Zhukov, L. (2022, May). Uncertainty estimation of transformer predictions for misclassification detection. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8237–8252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.566

Vazhentsev, A., Kuzmin, G., Tsvigun, A., Panchenko, A., Panov, M., Burtsev, M., & Shelmanov, A. (2023, July). Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11659–11681). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.652

Venkataramanan, A., Benbihi, A., Laviale, M., & Pradalier, C. (2023). Gaussian latent representations for uncertainty estimation using mahalanobis distance in deep classifiers, 4488–4497. Retrieved January 21, 2024, from https://openaccess.thecvf.com/content/ICCV2023W/UnCV/html/Venkataramanan_Gaussian_Latent_Representations_for_Uncertainty_Estimation_Using_Mahalanobis_Distance_in_ICCVW_2023_paper.html

Vincent, J. (2018, January 12). *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech* [The verge]. Retrieved July 5, 2023, from https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Wang, S., Zhou, T., & Bilmes, J. (2019, September). Jumpout : Improved dropout for deep neural networks with ReLUs. In K. Chaudhuri & R. Salakhut-

dinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 6668–6676, Vol. 97). PMLR. https://proceedings.mlr.press/v97/wang19q.html

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., & Li, Y. (2022). Mitigating neural network overconfidence with logit normalization [ISSN: 2640-3498]. *Proceedings of the 39th International Conference on Machine Learning*, 23631–23644. Retrieved April 4, 2023, from https://proceedings.mlr.press/v162/wei22d.html

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., & Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? [ISSN: 2640-3498]. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2282–2292. Retrieved August 23, 2023, from https://proceedings.mlr.press/v216/wimmer23a.html

Xiao, Y., & Wang, W. Y. (2019). Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 7322–7329. https://doi.org/10.1609/aaai.v33i01.33017322

Yu, J., Cristea, A. I., Harit, A., Sun, Z., Aduragba, O. T., Shi, L., & Moubayed, N. A. (2022). Efficient uncertainty quantification for multilabel text classification. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892871

Zakrzewski, C., De Vynck, G., Masih, N., & Mahtani, S. (2021, October 24). *How facebook neglected the rest of the world, fueling hate speech and violence in india* [Washington post] [Section: Tech Policy]. Retrieved June 1, 2023, from https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., & Wang, X. (2022a). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, *489*, 449–465. https://doi.org/https://doi.org/10.1016/j.neucom.2021.10.119

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., & Wang, X. (2022b). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, *489*, 449–465. https://doi.org/10.1016/j.neucom.2021.10.119

APPENDIX

Experiment Dissent:

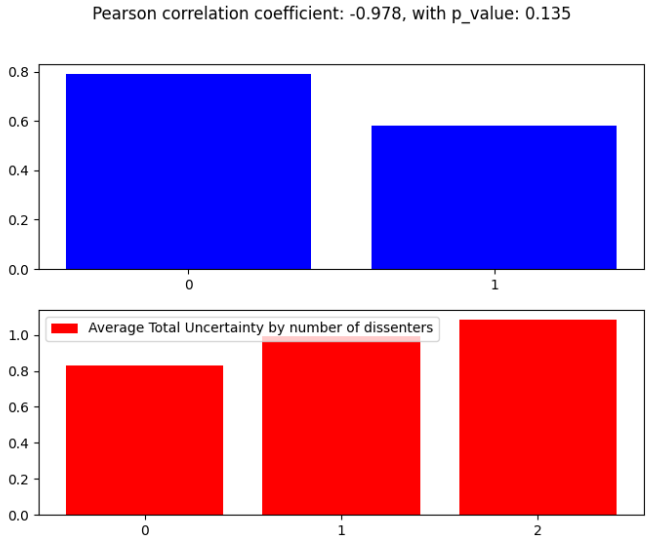

Figure 9: Total uncertainty shows the same correlation as aleatoric uncertainty in regard to the number of dissenters

Experiment target groups:

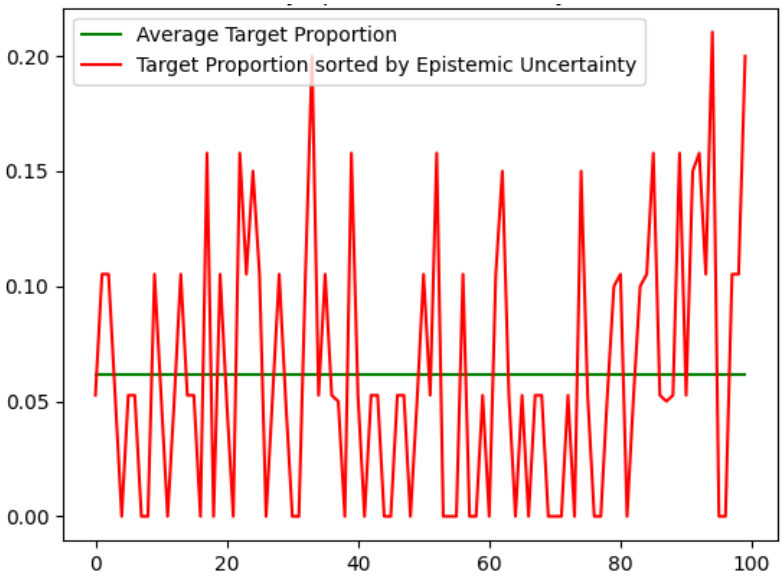

Figure 10: Target proportion grouped as percentage of test data, sorted by increasing epistemic uncertainty. Pearson correlation coefficient: 0.161 and p_value: 0.109

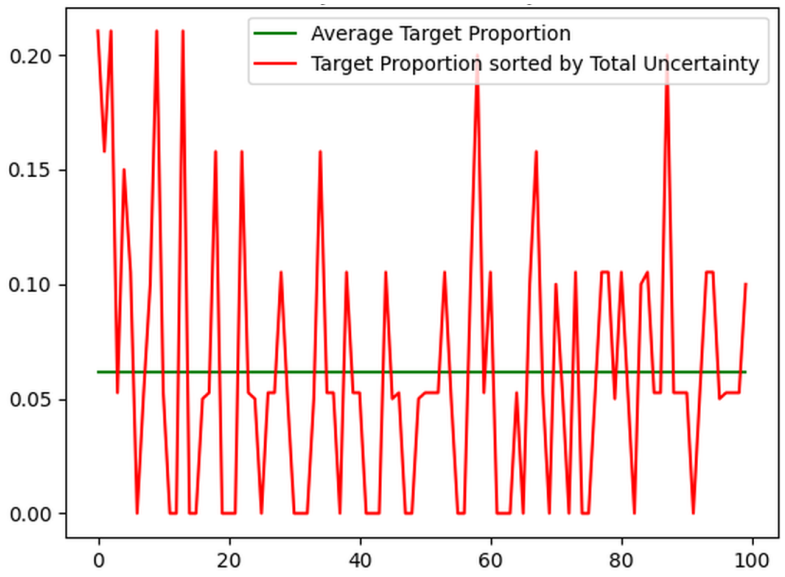

Figure 11: Target proportion grouped as percentage of test data, sorted by increasing total uncertainty. Pearson correlation coefficient: -0.035 and p_value: 0.732

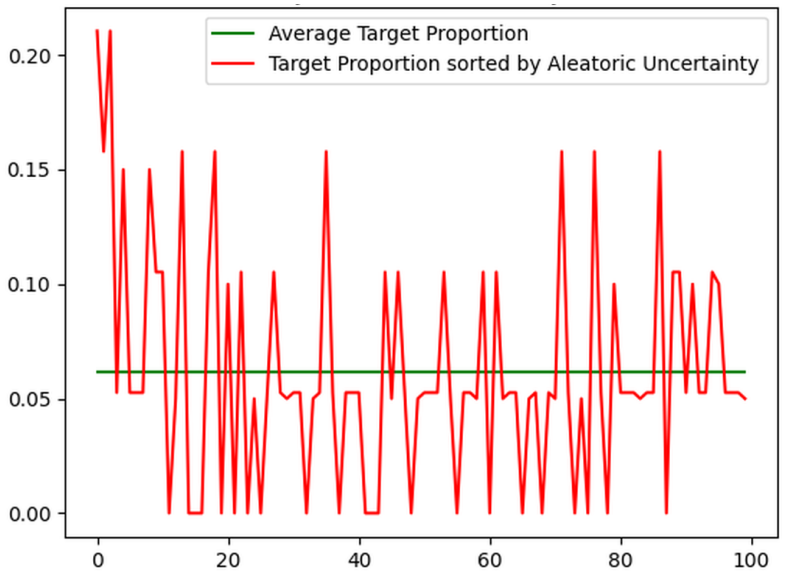

Figure 12: Target proportion grouped as percentage of test data, sorted by increasing aleatoric uncertainty. Pearson correlation coefficient: -0.079 and p_value: 0.433

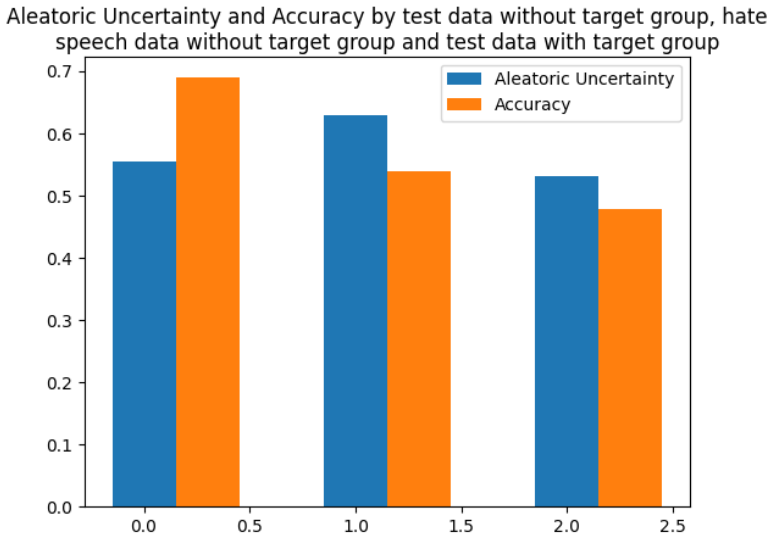

Figure 13: While accuracy decreases as expected for unknown target groups, aleatoric uncertainty does not increase compared to the baseline.

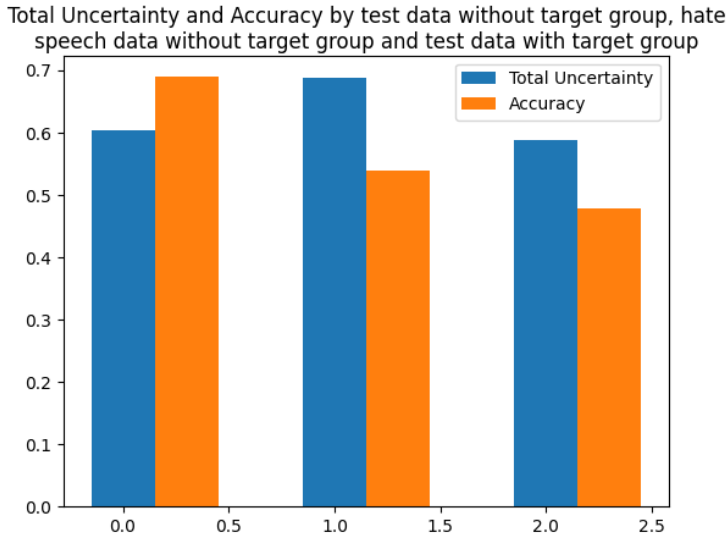

Figure 14: While accuracy decreases as expected for unknown target groups, total uncertainty does not increase compared to the baseline.

EHRENWÖRTLICHE ERKLÄRUNG

Hiermit wird versichert von Frank Walter die vorliegende MasterThesis gemäß §22 Abs. 7 APB der TU Darmstadt ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Mir ist bekannt, dass im Falle eines Plagiats (§38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Darmstadt, January 31, 2024

Frank Walter