Types of Out-of-Distribution Texts and How to Detect Them

Udit Arora[♠] **William Huang**[♣*] **He He**[♠]
[♠]New York University
[♣]Capital One
{uditarora,hhe}@nyu.edu, william.huang@capitalone.com

Abstract

Despite agreement on the importance of detecting out-of-distribution (OOD) examples, there is little consensus on the formal definition of OOD examples and how to best detect them. We categorize these examples by whether they exhibit a *background* shift or a *semantic* shift, and find that the two major approaches to OOD detection, model calibration and density estimation (language modeling for text), have distinct behavior on these types of OOD data. Across 14 pairs of in-distribution and OOD English natural language understanding datasets, we find that density estimation methods consistently beat calibration methods in background shift settings, while performing worse in semantic shift settings. In addition, we find that both methods generally fail to detect examples from challenge data, highlighting a weak spot for current methods. Since no single method works well across all settings, our results call for an explicit definition of OOD examples when evaluating different detection methods.

1 Introduction

Current NLP models work well when the training and test distributions are the same (e.g. from the same benchmark dataset). However, it is common to encounter out-of-distribution (OOD) examples that diverge from the training data once the model is deployed to real settings. When training and test distributions differ, current models tend to produce unreliable or even catastrophic predictions that hurt user trust (Ribeiro et al., 2020). Therefore, it is important to identify OOD inputs so that we can modify models' inference-time behavior by abstaining, asking for human feedback, or gathering additional information (Amodei et al., 2016).

Current work in NLP either focuses on specific tasks like intent classification in task-oriented dialogue (Zheng et al., 2020), or arbitrary in-distribution (ID) and OOD dataset pairs

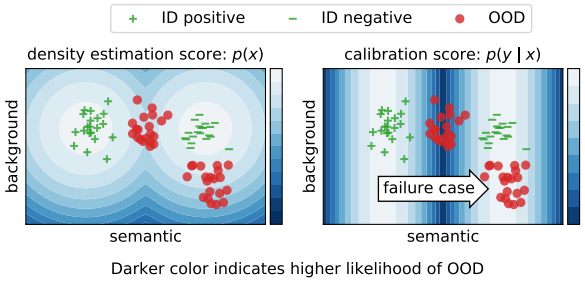

Figure 1: Illustration of semantic shift and background shift in $\mathbb{R}^2$. Each point consists of semantic features (x-axis) and background features (y-axis). OOD examples (red points) can shift in either direction. The background color indicates regions of ID (light) and OOD (dark) given by the density estimation method (left) and the calibration method (right). The calibration method fails to detect OOD examples due to background shift.

(Hendrycks et al., 2020b, 2019; Zhou and Chen, 2021), e.g. taking a sentiment classification dataset as ID and a natural language inference dataset as OOD. However, getting inputs intended for a different task is rare in realistic settings as users typically know the intended task. In practice, an example is considered OOD due to various reasons, e.g. being rare (Sagawa et al., 2020), out-of-domain (Daumé III, 2007), or adversarial (Carlini and Wagner, 2017). This broad range of distribution shifts makes it unreasonable to expect a detection algorithm to work well for arbitrary OOD examples without assumptions on the test distribution (Ahmed and Courville, 2020).

In this paper, we categorize OOD examples by common types of distribution shifts in NLP problems inspired by Ren et al. (2019) and Hsu et al. (2020). Specifically, we assume an input (e.g. a movie review) can be represented as background features (e.g. genre) that are invariant across different labels, and semantic features (e.g. sentiment words) that are discriminative for the prediction task. Correspondingly, at test time we consider two types of OOD examples characterized by a major

*Work done while at New York University.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10687–10701
November 7–11, 2021.

shift in the distribution of background and semantic features, respectively. While the two types of shifts often happen simultaneously, we note that there are realistic settings where distribution shift is dominated by one or the other. For example, background shift dominates when the domain or the style of the text changes (Pavlick and Tetreault, 2016), e.g. from news to tweets, and semantic shift dominates when unseen classes occur at test time, as in open-set classification (Scheirer et al., 2013).[1]

We use this categorization to evaluate two major approaches to OOD detection, namely calibration methods that use the model's prediction confidence (Hendrycks and Gimpel, 2017; Liang et al., 2018) and density estimation methods that fit a distribution of the training inputs (Nalisnick et al., 2019a; Winkens et al., 2020; Kirichenko et al., 2020). We show that the two approaches make implicit assumptions on the type of distribution shift, and result in behavioral differences under each type of shift. By studying ID/OOD pairs constructed from both simulations and real datasets, we find that the density estimation method better accounts for shifts in background features, consistently outperforming the calibration method on *background* shift pairs. We further see the opposite in *semantic* shift pairs, with the calibration method consistently yielding higher performance.

In addition, we analyze the detection performance on challenge datasets (McCoy et al., 2019a; Naik et al., 2018b) through the lens of background/semantic shift. We find that these challenge datasets provide interesting failure cases for both methods. Calibration methods completely fail when the model is over-confident due to spurious semantic features. While density estimation methods are slightly more robust, language models are easily fooled by repetitions that significantly increase the probability of a piece of text. Together, our findings suggest that better definitions of OOD and corresponding evaluation datasets are required for both model development and fair comparison of OOD detection methods.

[1]We exclude *task* shift where the OOD examples are from a different task, e.g. textual entailment inputs for a text classification model, because it is less likely to happen in realistic settings where users are often aware of the intended use of the model.

2 Categorization of OOD Examples

2.1 Problem Statement

Consider classification tasks where each example consists of an input $x \in \mathcal{X}$ and its label $y \in \mathcal{Y}$. In the task of OOD detection, we are given a training dataset $\mathcal{D}_{\text{train}}$ of (x, y) pairs sampled from the training data distribution $p(x, y)$. At inference time, given an input $x' \in \mathcal{X}$ the goal of OOD detection is to identify whether x' is a sample drawn from $p(x, y)$.

2.2 Types of Distribution Shifts

As in (Ren et al., 2019), we assume that any representation of the input x, $\phi(x)$, can be decomposed into two independent and disjoint components: the background features $\phi_b(x) \in \mathbb{R}^m$ and the semantic features $\phi_s(x) \in \mathbb{R}^n$. Formally, we have

$$\phi(x) = [\phi_s(x); \phi_b(x)], \tag{1}$$
$$p(x) = p(\phi_s(x))p(\phi_b(x)) \tag{2}$$

Further, we assume that $\phi_b(x)$ is independent of the label while $\phi_s(x)$ is not. Formally, $\forall y \in \mathcal{Y}$,

$$p(\phi_b(x) \mid y) = p(\phi_b(x)), \tag{3}$$
$$p(\phi_s(x) \mid y) \neq p(\phi_s(x)) \tag{4}$$

Note that p refers to the ground truth distribution, as opposed to one learned by a model.

Intuitively, the background features consist of population-level statistics that do not depend on the label, whereas the semantic features have a strong correlation with the label. A similar decomposition is also used in previous work on style transfer (Fu et al., 2018), where a sentence is decomposed into the content (semantic) and style (background) representations in the embedding space.

Based on this decomposition, we classify the types of OOD data as either *semantic* or *background* shift based on whether the distribution shift is driven by changes in $\phi_s(x)$ or $\phi_b(x)$, respectively. An example of background shift is a sentiment classification corpus with reviews from IMDB versus GoodReads where phrases indicating positive reviews (e.g. "best", "beautifully") are roughly the same while the background phrases change significantly (e.g. "movie" vs "book"). On the other hand, semantic shift happens when we encounter unseen classes at test time, e.g. a dialogue system for booking flight tickets receiving a request for meal vouchers (Zheng et al., 2020), or a

question-answering system handling unanswerable questions (Rajpurkar et al., 2018). We note that the two types of shifts may happen simultaneously in the real world, and our categorization is based on the most prominent type of shift.

3 OOD Detection Methods

To classify an input $x \in \mathcal{X}$ as ID or OOD, we produce a score $s(x)$ and classify it as OOD if $s(x) < \gamma$, where γ is a pre-defined threshold. Most methods differ by how they define $s(x)$. Below we describe two types of methods commonly used for OOD detection.

Calibration methods. These methods use the model's prediction confidence as the score. A well-calibrated model's confidence score reflects the likelihood of the predicted label being correct. Since the performance on OOD data is usually lower than on ID data, lower confidence suggests that the input is more likely to be OOD. The simplest method to obtain the confidence score is to directly use the conditional probability produced by a probabilistic classifier p_{model}, referred to as maximum softmax probability (*MSP*; Hendrycks and Gimpel, 2017). Formally,

$$s_{MSP}(x) = \max_{k \in \mathcal{Y}} p_{\text{model}}(y = k \mid x). \quad (5)$$

While there exist more sophisticated methods that take additional calibration steps (Liang et al., 2018; Lee et al., 2018), *MSP* proves to be a strong baseline, especially when p_{model} is fine-tuned from pre-trained transformers (Hendrycks et al., 2020b; Desai and Durrett, 2020).

Density estimation methods. These methods use the likelihood of the input given by a density estimator as the score. For text or sequence data, a language model p_{LM} is typically used to estimate $p(x)$ (Ren et al., 2019). To avoid bias due to the length of the sequence (see analysis in Appendix A), we use the token perplexity (*PPL*) as the score. Formally, given a sequence $x = (x_1, \ldots, x_T)$,

$$s_{PPL}(x) = \exp\left\{\frac{1}{T}\sum_{t=1}^{T} \log p_{\text{LM}}(x_t \mid x_{1:t-1})\right\} \quad (6)$$

While there are many works on density estimation methods using flow-based models in computer vision (e.g. Nalisnick et al., 2019a; Zhang et al., 2020a), there is limited work experimenting with density estimation methods for OOD detection on text (Lee et al., 2020).

Implicit assumptions on OOD. One key question in OOD detection is how the distribution shifts at test time, i.e. what characterizes the difference between ID and OOD examples. Without access to OOD data during training, the knowledge must be incorporated into the detector through some inductive bias. Calibration methods rely on $p(y \mid x)$ estimated by a classifier, thus they are more influenced by the semantic features which are correlated with the label. We can see this formally by

$$\begin{aligned} p(y \mid x) &\propto p(x \mid y)p(y) && (7)\\ &= p(\phi_b(x) \mid y)p(\phi_s(x) \mid y)p(y) && (8)\\ &\propto p(\phi_s(x) \mid y)p(y). && (9) \end{aligned}$$

In contrast, density estimation methods are sensitive to all components of the input, including both background and semantic features, even in situations where distribution shifts are predominately driven by one particular type. In the following sections, we examine how these implicit assumptions impact performance on different ID/OOD pairs.

4 Simulation of Distribution Shifts

As an illustrative example, we construct a toy OOD detection problem using a binary classification setting similar to the one depicted in Figure 1. This allows us to remove estimation errors and study optimal calibration and density estimation detectors under controlled semantic and background shifts.

4.1 Data Generation

We generate the ID examples from a Gaussian Mixture Model (GMM):

$$y = \begin{cases} 0 & \text{w.p. } 0.5 \\ 1 & \text{otherwise} \end{cases}, \quad (10)$$

$$x \mid y = i \sim \mathcal{N}(\mu^i, \Sigma). \quad (11)$$

The centroids are sets of semantic and background features such that $\mu^1 = [\mu_s, \mu_b]$ and $\mu^0 = [-\mu_s, \mu_b]$, where $\mu_s \in \mathbb{R}^n$ and $\mu_b \in \mathbb{R}^m$. In the 2D case in Figure 1, this corresponds to the two Gaussian clusters where the first component is the semantic feature and the second is the background feature.

In this case, we know the true calibrated score $p(y \mid x)$ and the true density $p(x)$ given any inputs.

Specifically, the optimal classifier is given by the Linear Discriminant Analysis (LDA) predictor. By setting Σ to the identity matrix, it corresponds to a linear classifier with weights $[2\mu_s, \mathbf{0}_b]$, where $\mathbf{0}_b \in \mathbb{R}^m$ is a vector of all 0s. For simplicity, we set $\mu_s = \mathbf{1}_s$ and $\mu_b = \mathbf{0}_b$, where $\mathbf{1}_s \in \mathbb{R}^n, \mathbf{0}_b \in \mathbb{R}^m$ are vectors of all 0s.

4.2 Semantic Shift

We generate sets of OOD examples using a semantic shift by varying the overlap of ID and OOD semantic features. Formally, we vary the overlap rate r such that

$$r = \frac{|\mu_s \cap \mu_s^{\text{Shift}}|}{|\mu_s|} \tag{12}$$

where $\mu_s, \mu_s^{\text{Shift}} \in \mathbb{R}^n$ are the set of semantic features for ID and OOD, respectively, $\mu_s \cap \mu_s^{\text{Shift}}$ represents the common features between the two, and $|\cdot|$ denotes the number of elements.

We fix the total dimensions to $n + m = 200$ and set $n = 40$ (semantic features) and $m = 160$ (background features). Further, we vary r by increments of 10%. Larger r indicates stronger semantic shift. For each r, we randomly sample ID and OOD semantic features and report the mean over 20 trials with 95% confidence bands in Figure 2.

4.3 Background Shift

We generate sets of OOD examples using a background shift by applying a displacement vector $z = [\mathbf{0}_s, z_b]$ to the two means. Formally,

$$\mu^{i,\,\text{Shift}} = \mu^i + z \tag{13}$$

where $\mathbf{0}_s \in \mathbb{R}^n$ is a vector of all 0s.

We set $z = \alpha[\mathbf{0}_s, \mathbf{1}_b]$, where $\mathbf{1}_b \in \mathbb{R}^m$ is a vector of 1s. Note that this shift corresponds to a translation of the ID distribution along the direction of μ_b. We set the total dimensions to $n + m = 200$ while varying the split between semantic (n) and background (m) components by increments of 20.

4.4 Simulation Results

Figure 2 shows the OOD detection performance of our simulated experiment. We use Area Under the Receiver Operating Characteristics (AUROC) as our performance metric.

We see that the calibration method generally outperforms density estimation. Further, the performance gap between the two methods decreases as both methods approach near-perfect performance

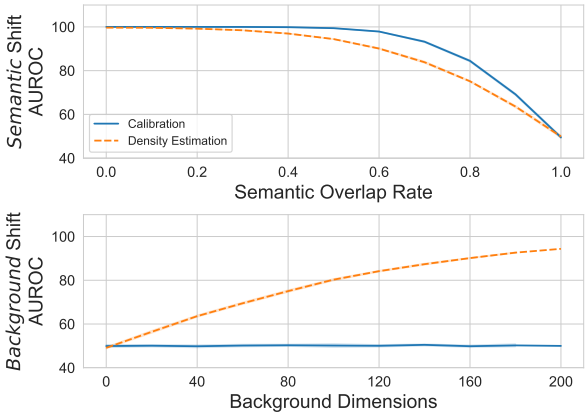

Figure 2: Area Under Receiver Operating Characteristics (AUROC) of calibration (blue) and density estimation (orange) methods for OOD detection using our toy binary classification problem. The calibration method outperforms the density estimation method under larger semantic shifts while the opposite is true under larger background shifts.

under large semantic shifts with no overlap in semantic features, and approach chance under no semantic shift with completely overlapping semantic features. However, the calibration method is unable to improve performance under background shifts in either regime because the background features do not contribute to $p(y \mid x)$ as the LDA weights are 0 for these components (Section 4.1). We find these results in line with our expectations and use them to drive our intuition when evaluating both types of OOD detection methods for real text data.

5 Experiments and Analysis

We perform head-to-head comparisons of calibration and density estimation methods on 14 ID/OOD pairs categorized as either background shift or semantic shift, as well as 8 pairs from challenge datasets.

5.1 Setup

OOD detectors. Recall that the **calibration method** *MSP* relies on a classifier trained on the ID data. We fine-tune the RoBERTa (Liu et al., 2019) model on the ID data and compute its prediction probabilities (see Equation (5)). For the **density estimation** method *PPL*, we fine-tune GPT-2 (Radford et al., 2019) on the ID data and use perplexity as the OOD score (see Equation (6)).[2] To control for model size of the two methods, we choose

[2]We also use the sentence probability ($p(x)$) as the score, but find it highly sensitive to sequence lengths (Appendix A).

RoBERTa$_{\text{Base}}$ and GPT-2$_{\text{Small}}$, which have 110M and 117M parameters, respectively. We also experiment with two larger models, RoBERTa$_{\text{Large}}$ and GPT-2$_{\text{Medium}}$ with 355M and 345M parameters, respectively.

We evaluate the OOD detectors by AUROC and the False Alarm Rate at 95% Recall (FAR95), which measures the misclassification rate of ID examples at 95% OOD recall. Both metrics show similar trends (see Appendix B for FAR95 results).

Training details. For RoBERTa, we fine-tune the model for 3 epochs on the training split of ID data with a learning rate of 1e-5 and a batch size of 16. For GPT-2, we fine-tune the model for 1 epoch on the training split of ID data for the language modeling task, using a learning rate of 5e-5 and a batch size of 8. [3]

Oracle detectors. To get an estimate of the upper bound of OOD detection performance, we consider the situation where we have access to the OOD data and can directly learn an OOD classifier. Specifically, we train a logistic regression model with bag-of-words features using 80% of the test data and report results on the remaining 20%.

5.2 Semantic Shift

Recall that the distribution of discriminative features changes in the semantic shift setting, i.e. $p_{\text{train}}(\phi_s(x)) \neq p_{\text{test}}(\phi_s(x))$ (Section 2). We create semantic shift pairs by including test examples from classes unseen during training. Thus, semantic features useful for classifying the training data are not representative in the test set.

We use the News Category (Misra, 2018) and DBPedia Ontology Classification (Zhang et al., 2015) multiclass classification datasets to create two ID/OOD pairs. The News Category dataset consists of HuffPost news data. We use the examples from the five most frequent classes as ID (News Top-5) and the data from the remaining 36 classes as OOD (News Rest). The DBPedia Ontology Classification dataset consists of data from Wikipedia extracted from 14 non-overlapping classes of DBPedia 2014 (Lehmann et al., 2015). We use examples from the first four classes by class number as ID (DBPedia Top-4) and the rest as OOD (DBPedia Rest).

[3]Our code can be found at https://github.com/uditarora/ood-text-emnlp.

ID	OOD	AUROC		
		PPL	*MSP*	Oracle
News Top-5	News Rest	60.2	**78.9**	72.0
DBPedia Top-4	DBPedia Rest	75.4	**88.8**	99.6

Table 1: Performance on semantic shifts, with higher score (among *PPL*/*MSP*) in **bold**. We can see that the calibration method using *MSP* significantly outperforms the density estimation methods.

Results. Table 1 shows the results for our semantic shift pairs. The calibration method consistently outperforms the density estimation method, indicating that calibration methods are better suited for scenarios with large semantic shifts, which is in line with our simulation results (Section 4).

5.3 Background Shift

Recall that background features (e.g. formality) do not depend on the label. Therefore, we consider domain shift in sentiment classification and natural language inference (NLI) datasets.

For our analysis, we use the SST-2 (Socher et al., 2013), IMDB (Maas et al., 2011), and Yelp Polarity (Zhang et al., 2015) binary sentiment classification datasets. The SST-2 and IMDB datasets consist of movie reviews with different lengths. Meanwhile, the Yelp polarity dataset contains reviews for different businesses, representing a domain shift from SST-2 and IMDB. Each of these datasets is used as ID/OOD, using the validation split of SST-2 and test split of IMDB and Yelp Polarity for evaluation.

We also use the SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018) and RTE (from GLUE, Wang et al., 2018a) datasets. SNLI and MNLI consist of NLI examples sourced from different genres. RTE comprises of examples sourced from a different domain. Where there is some change in semantic information since the task has the two labels (*entailment* and *non-entailment*) as opposed to three (*entailment*, *neutral* and *contradiction*) in SNLI and MNLI,[4] domain/background shift is more prominent since the semantic features for the NLI task are similar. Each of these datasets is used as either ID or OOD, and we use the validation set of the OOD data for evaluation.

Results. Table 2 shows the results for binary sentiment classification and NLI domain shifts.

[4]Both neutral and contradiction are considered as non-entailment when evaluating accuracy with RTE vs SNLI/MNLI or vice-versa.

ID	OOD	AUROC			Accuracy	
		PPL	*MSP*	Oracle	OOD (Δ)	ID
SST-2	IMDB	**97.9**	66.2	100.0	92.0 (-1.8)	93.8
	Yelp	**98.7**	57.5	99.8	94.4 (+0.6)	
IMDB	SST-2	**96.9**	82.6	100.0	89.2 (-6.3)	95.5
	Yelp	**77.9**	67.1	100.0	95.4 (-0.1)	
Yelp	SST-2	**98.9**	85.9	99.8	88.9 (-9.3)	98.2
	IMDB	**86.6**	61.8	100.0	93.2 (-5.0)	
SNLI	RTE	**94.6**	78.7	99.8	67.5 (-22.6)	90.1
	MNLI	**96.7**	75.6	99.7	77.9 (-12.2)	
RTE	SNLI	**81.2**	45.1	99.7	82.0 (+6.9)	75.1
	MNLI	**81.4**	55.5	97.0	77.3 (+2.2)	
MNLI	SNLI	**75.7**	56.1	99.7	80.4 (-4.4)	84.8
	RTE	68.0	**76.5**	96.7	76.5 (-8.3)	

Table 2: Performance on background shifts caused by shift in domain. For each pair, higher score obtained (by *PPL* or *MSP*) is in **bold**. The density estimation method using *PPL* outperforms the calibration method.

The density estimation method consistently outperforms the calibration method (for all pairs except MNLI vs RTE), indicating that *PPL* is more sensitive to changes in background features. Further, in cases where the discriminative model generalizes well (as evident by the small difference in ID and OOD accuracy numbers), we find that the calibration method performance is close to random (50) because a well-calibrated model also has higher confidence on its correct OOD predictions.

We note that the discriminative models tend to generalize well here, hence it might be better to focus on domain adaptation instead of OOD detection when the shift is predominantly a background shift. We discuss this further in Section 6.

5.4 Analysis

Controlled distribution shifts. We use two controlled distribution shift experiments on real text data to further study the framework of semantic and background shifts. For background shift, we append different amounts of text from Wikitext (Merity et al., 2017) and Civil Comments (Borkan et al., 2019a) to SST-2 examples to create synthetic ID and OOD examples, respectively. We append the unrelated texts with lengths $\in (25, 50, 100, 150, 200)$ words. For semantic shift, we use the News Category dataset and move classes from ID to OOD. We start with the top 40 ID classes by frequency and move classes in increments of 10. The ID coverage of semantic information decreases as more classes move to the OOD subset, resulting in a larger semantic shift.

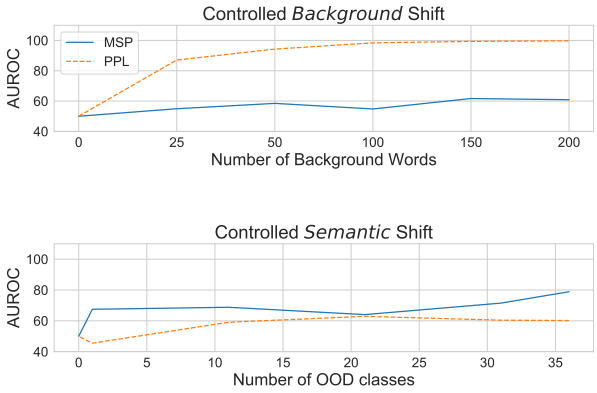

Figure 3: AUROC of *PPL* (orange) and *MSP* (blue) for controlled background and semantic shift experiments. The density estimation method performance improves as we increase the amount of background shift by appending longer texts, and the calibration method performance increases as we increase the amount of semantic shift by moving more classes to OOD.

ID	OOD	Base		Large		
		PPL	*MSP*	*PPL*	*MSP*	Oracle
IMDB	Yelp	**77.9**	67.1	75.5	74.5	100.0
News Top-5	News Rest	60.2	78.9	61.7	**79.1**	72.0

Table 3: Performance of Base and Large models for a background shift pair and semantic shift pair each, with higher score in **bold**. The larger discriminative model helps close the performance gap between the calibration method and density estimation method for background shift.

Results. Figure 3 shows the AUROC score obtained from both methods for our controlled distribution shift experiments. We see that the density estimation method is more sensitive to the amount of synthetic background text than calibration methods, and that the calibration method is more sensitive to the number of ID/OOD classes. This is in line with our intuition about the shifts and the results we obtain from simulated data (Section 4).

Larger models. Table 3 shows the results using larger models for OOD detection. We observe that the larger discriminative model achieves a much higher score for the background shift pair, closing the gap with the language model performance. We speculate that the larger model is able to learn some of the background features in its representation. The performance for the semantic shift pair is largely unchanged when using the larger models.

5.5 Challenge Data

Challenge datasets are designed to target either superficial heuristics adopted by a model (e.g. premise-hypothesis overlap) or model deficiencies (e.g. numerical reasoning in NLI), which creates significant challenges for deployed models (Ribeiro et al., 2020). It is therefore desirable to abstain on detected OOD examples. We consider the following challenge datasets.

Human-generated challenge data. Kaushik et al. (2020) crowdsourced a set of counterfactually-augmented IMDB examples (c-IMDB) by instructing annotators to minimally edit examples to yield counterfactual labels. This changes the distribution of semantic features with high correlation to labels such that $p_{\text{train}}(\phi_s(x)) \neq p_{\text{test}}(\phi_s(x))$, creating a semantic shift. We consider IMDB as ID and c-IMDB as OOD, combining the training, validation, and test splits of c-IMDB for evaluation.

Rule-based challenge data. HANS (McCoy et al., 2019b) consists of template-based examples that have high premise-hypothesis overlap but are non-entailment, which mainly results in background shift due to the specific templates/syntax. Similarly, the Stress Test dataset (Naik et al., 2018a) is a set of automatically generated examples designed to evaluate common errors from NLI models. We categorize the type of distribution shifts from these test categories with respect to MNLI (ID) depending on whether they append "background" phrases to the ID examples or replace discriminative phrases (Table 4).

Antonym (changing premise to obtain an antonymous hypothesis resulting in contradiction despite high overlap) and Numerical Reasoning (different semantic information than MNLI training set) constitute semantic shifts, as the set of semantic features now focus on specific types of entailment reasoning (e.g. antonymy and numerical representation). Negation (appending "and false is not true" to hypothesis), Spelling Errors (randomly introducing spelling errors in one premise word), Word Overlap (appending "and true is true" to each hypothesis), and Length Mismatch (appending a repetitive phrase "and true is true" five times to the premise) constitute background shifts because they introduce population level changes (e.g. appending "and true is true" to each hypothesis) that are unrelated to the entailment conditions of each example.

ID	OOD	Shift	AUROC *PPL*	AUROC *MSP*	AUROC Oracle
IMDB	c-IMDB	Semantic	53.5	**63.7**	77.5
MNLI	HANS	Background	**98.3**	55.0	100.0
	Negation	Background	44.5	**60.5**	99.9
	Len. Mismatch	Background	19.6	**51.6**	100.0
	Spell. Error	Background	43.9	**57.7**	98.4
	Word Overlap	Background	42.4	**61.7**	99.8
	Antonym	Semantic	4.5	**55.3**	97.3
	Num. Reason.	Semantic	27.5	**75.8**	99.7

Table 4: AUROC scores obtained using *PPL*, *MSP* and Oracle for challenge data. The primary type of shift observed is indicated in the 'Shift' column. Higher performance (among *MSP*/*PPL*) for each pair is in **bold**. We can see that both methods struggle with most types of challenge data.

We consider the matched Negation, Spelling Errors, Word Overlap and Length Mismatch examples from the Stress Test as background shifts, and the Numerical Reasoning and Antonym examples as semantic shifts. We consider MNLI as ID for these challenge examples and use the validation split of HANS and MNLI for evaluation.

Failure case 1: spurious semantic features. Challenge data is often constructed to target *spurious features* (e.g. premise-hypothesis overlap for NLI) that are useful on the training set but do not correlate with the label in general, e.g. on the test set. Therefore, a discriminative model would be *over-confident* on the OOD examples because the spurious semantic features that were discriminative during training, while still prominent, are no longer predictive of the label. As a result, in Table 4, *MSP* struggles with most challenge data, achieving an AUROC score close to random (50). On the other hand, the density estimation method achieves almost perfect performance on HANS.

Failure case 2: small shifts. While density estimation methods perform better in background shift settings, our simulation results show that they still struggle to detect small shifts when the ID and OOD distributions largely overlap. Table 4 shows similar findings for Negation and Word Overlap Stress Test categories that append short phrases (e.g. "and true is true") to each ID hypothesis.

Failure case 3: repetition. For Antonym, Numerical Reasoning, and Length Mismatch, *PPL* performance is *significantly worse than random*, indicating that our language model assigns higher likelihoods to OOD than ID examples. These challenge

examples contain highly repetitive phrases (e.g. appending "*and true is true*" five times in Length Mismatch, or high overlap between premise and hypothesis in Numerical Reasoning and Antonym), which is known to yield high likelihood under recursive language models (Holtzman et al., 2020). Thus repetition may be used as an attack to language model-based OOD detectors.

Overall, the performance of both methods drops significantly on the challenge datasets. Among these, human-generated counterfactual data is the most difficult to detect, and rule-based challenge data can contain unnatural patterns that cause unexpected behavior.

5.6 Discussion

The performance of calibration and density estimation methods on OOD examples categorized along the lines of semantic and background shift provides us with insights that can be useful in improving OOD detection. This framework can be used to build better evaluation benchmarks that focus on different challenges in OOD detection. A choice between the two methods can also be made based on the anticipated distribution shift at test time, i.e, using calibration methods when detecting semantic shift is more important, and using density estimation methods to detect background shifts. However, we observe failure cases from challenge examples, with density estimation methods failing to detect OOD examples with repetition and small shifts, and calibration methods failing to detect most challenge examples. This indicates that these challenge examples constitute a type of OOD that target the weaknesses of both approaches. This highlights the room for a more explicit definition of OOD to progress the development of OOD detection methods and create benchmarks that reflect realistic distribution shifts.

6 Related Work

Distribution shift in the wild. Most early works on OOD detection make no distinctions on the type of distribution shift observed at test time, and create synthetic ID/OOD pairs using different datasets based on the setup in Hendrycks and Gimpel (2017). Recently, there is an increasing interest in studying real-world distribution shifts (Ahmed and Courville, 2020; Hsu et al., 2020; Hendrycks et al., 2020a; Koh et al., 2020a). On these benchmarks with a diverse set of distribution shifts, no single detection method wins across the board. We explore the framework of characterization of distribution shifts along the two axes of semantic shift and background (or non-semantic) shift, shedding light on the performance of current methods.

OOD detection in NLP. Even though OOD detection is crucial in production (e.g. dialogue systems (Ryu et al., 2018)) and high-stake applications (e.g. healthcare (Borjali et al., 2020)), it has received relatively less attention in NLP until recently. Recent works evaluated/improved the calibration of pretrained transformer models (Hendrycks et al., 2020b; Goyal and Durrett, 2020; Kong et al., 2020; Zhou and Chen, 2021). They show that while pretrained transformers are better calibrated, making them better at detecting OOD data than previous models, there is scope for improvement. Our analysis reveals one limitation of calibration-based detection when faced with a background shift. Other works focus on specific tasks, including prototypical network for low-resource text classification (Tan et al., 2019) and data augmentation for intent classification (Zheng et al., 2020).

Inductive bias in OOD detection. Our work shows that the effectiveness of a method largely depends on whether its assumption on the distribution shift matches the test data. One straightforward way to incorporate prior knowledge on the type of distribution shift is through augmenting similar OOD data during training, i.e., the so-called outlier exposure method (Hendrycks et al., 2019), which has been shown to be effective on question answering (Kamath et al., 2020). Given that the right type of OOD data can be difficult to obtain, another line of work uses a hybrid of calibration and density estimation methods to achieve a balance between capturing semantic features and background features. These models are usually trained with both a discriminative loss and a generative (or self-supervised) loss (Winkens et al., 2020; Zhang et al., 2020a; Nalisnick et al., 2019b).

Domain adaptation versus OOD detection. There are two ways of handling the effect of OOD data: 1) build models that perform well across domains (i.e., background shifts), i.e., domain adaptation (Chu and Wang, 2018; Kashyap et al., 2021) or 2) allow models to detect a shift in data distribution, and potentially abstain from making a prediction. In our setting (2), we want to guard

against all types of OOD data without any access to it, unlike domain adaptation which usually relies on access to OOD data. This setting can be more important than (1) for safety-critical applications, such as those in healthcare, because the potential cost of an incorrect prediction is greater, motivating a more conservative approach to handling OOD data by abstaining. This could also help improve performance in selective prediction (Kamath et al., 2020; Xin et al., 2021).

7 Conclusion

Despite the extensive literature on outlier and OOD detection, previous work in NLP tends to lack consensus on a rigorous definition of OOD examples, instead relying on arbitrary dataset pairs from different tasks. In our work, we approach this problem in natural text and simulated data by categorizing OOD examples as either *background* or *semantic* shifts and study the performance of two common OOD detection methods—calibration and density estimation. For both types of data, we find that density estimation methods outperform calibration methods under background shifts while the opposite is true under semantic shifts. However, we find several failure cases from challenge examples that target model shortcomings.

As explained in Section 2, we assume that ϕ_s and ϕ_b map x to two disjoint sets of components for simplicity. This assumption helps us simplify the framework and compare the two types of detection methods in relation to the two types of shifts. While this simplified framework explains much of the differences between the two methods, failure cases from challenge examples highlight the room for better frameworks and a more explicit definition of OOD to progress the development of OOD detection methods. Such a definition can inform the creation of benchmarks on OOD detection that reflect realistic distribution shifts.

Defining (or at least explicitly stating) the types of OOD examples that predictors are designed to target can also guide future modeling decisions between using calibration and density estimation methods, and help improve detection. Some promising directions include test-time fine-tuning (Sun et al., 2020) and data augmentation (Chen et al., 2020), which can be guided towards a specific type of distribution shift for improved detection performance against it. Finally, the methods we studied work well for one type of shift, which motivates the use of hybrid models (Zhang et al., 2020b; Liu and Abbeel, 2020) that use both calibration and density estimation when both types of shift occur at the same time.

Ethical Considerations

As society continues to rely on automated machine learning systems to make important decisions that affect human lives, OOD detection becomes increasingly vital to ensure that these systems can detect natural shifts in domain and semantics. If medical chat-bots cannot recognize that new disease variants or rare co-morbidities are OOD while diagnosing patients, they will likely provide faulty and potentially harmful recommendations [5] if they don't contextualize their uncertainty. We believe that implementing OOD detection, especially for more challenging but commonly occurring semantic shifts should be part of any long-lasting production model.

In addition, OOD detection can be used to identify and alter model behavior when encountering data related to minority groups. For example, Koh et al. (2020b) present a modified version of the CivilComments dataset (Borkan et al., 2019b), with the task of identifying toxic user comments on online platforms. They consider domain annotations for each comment based on whether the comment mentions each of 8 demographic identities - *male, female, LGBTQ, Christian, Muslim, other religions, Black and White*. They note that a standard BERT-based model trained using ERM performs poorly on the worst group, with a 34.2 % drop in accuracy as compared to the average. Such models may lead to unintended consequences like flagging a comment as toxic just because it mentions some demographic identities, or in other words, belongs to some domains. Our work can be useful in altering the inference-time behavior of such models upon detection of such domains which constitute a larger degree of background shift. Of course, nefarious agents could use the same pipeline to alter model behavior to identify and discriminate against demographics that display such background shifts.

Acknowledgements

We thank the anonymous reviewers, Ethan Perez, Angelica Chen and other members of the Machine Learning for Language Lab at New York University for their thoughtful suggestions on improving

[5] https://www.nabla.com/blog/gpt-3/

the paper. We also want to thank Diksha Meghwal, Vaibhav Gadodia and Ambuj Ojha for their help with an initial version of the project and experimentation setup.

References

Faruk Ahmed and Aaron C. Courville. 2020. Detecting semantic anomalies. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3154–3162. AAAI Press.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *CoRR*, abs/1606.06565.

Alireza Borjali, Martin Magneli, David Shin, Henrik Malchau, Orhun K. Muratoglu, and Kartik M. Varadarajan. 2020. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. *CoRR*, abs/2004.08333.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019a. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019b. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Nicholas Carlini and David A. Wagner. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 3–14. ACM.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2020. Robust out-of-distribution detection via informative outlier mining. *CoRR*, abs/2006.15207.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1304–1319. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 3592–3603. Association for Computational Linguistics.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2020a. Scaling out-of-distribution detection for real-world settings. *CoRR*, abs/1911.11132.

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.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020b. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2744–2751. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier

exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2020. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10948–10957. IEEE.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5684–5696. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: A survey and empirical analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1830–1849. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2020. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020a. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020b. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1326–1340. Association for Computational Linguistics.

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*, pages 7167–7177.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity. *CoRR*, abs/2006.04666.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Hao Liu and Pieter Abbeel. 2020. Hybrid discriminative-generative training via contrastive learning. *CoRR*, abs/2007.09070.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019a. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the wrong reasons: Diagnosing syntactic

heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Rishabh Misra. 2018. News category dataset.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018a. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig. 2018b. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2340–2353. Association for Computational Linguistics.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. 2019a. Do deep generative models know what they don't know? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. 2019b. Hybrid models with deep and invertible features. 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 4723–4732. PMLR.

Ellie Pavlick and Joel R. Tetreault. 2016. An empirical analysis of formality in online communication. *Trans. Assoc. Comput. Linguistics*, 4:61–74.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners.

P. Rajpurkar, R. Jia, and P. Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32, pages 14707–14718. Curran Associates, Inc.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4902–4912. Association for Computational Linguistics.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 714–718. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1757–1772.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. 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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3564–3570. Association for Computational Linguistics.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman.

2018b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, A. Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. 2020. Contrastive training for improved out-of-distribution detection. *CoRR*, abs/2007.05566.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1040–1051. Association for Computational Linguistics.

Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. 2020a. Hybrid models for open set recognition. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 102–117. Springer.

Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. 2020b. Hybrid models for open set recognition. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 102–117. Springer.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1198–1209.

Wenxuan Zhou and Muhao Chen. 2021. Contrastive out-of-distribution detection for pretrained transformers. *CoRR*, abs/2104.08812.

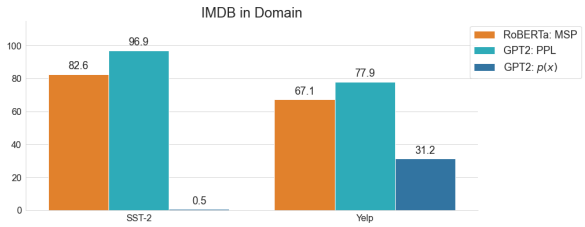

Figure 4: OOD detection performance as measured by AUROC using different measures for binary sentiment classification based background shift, using IMDB as ID data. We can see that using $\log p(x)$ as a measure is highly noisy due to its dependency on sequence lengths.

ID	OOD	FAR95 (↓) *PPL*	*MSP*	Oracle
News Top-5	News Rest	88.5	**75.7**	80.4
DBPedia Top-4	DBPedia Rest	**78.3**	86.3	1.3

Table 5: FAR95 scores obtained using *PPL*, *MSP* and Oracle for semantic shifts, with lower score (among *PPL*/*MSP*) in **bold**.

ID	OOD	FAR95 (↓) *PPL*	*MSP*	Oracle
SST-2	IMDB	**8.6**	76.5	0.0
	Yelp	**5.2**	83.0	0.0
IMDB	SST-2	**17.0**	47.7	0.2
	Yelp	**70.2**	82.6	0.0
Yelp	SST-2	**3.1**	45.4	1.1
	IMDB	**36.2**	90.4	0.0
SNLI	RTE	**19.1**	61.4	0.7
	MNLI	**14.7**	62.5	0.3
RTE	SNLI	**62.5**	95.3	0.0
	MNLI	**64.3**	93.9	10.3
MNLI	SNLI	**70.9**	84.6	1.2
	RTE	93.2	**69.8**	6.2

Table 6: FAR95 scores obtained using *PPL*, *MSP* and Oracle for background shift caused by shift in domain. For each pair, lower score obtained (by *PPL* or *MSP*) is in **bold**.

A Example Probability

We additionally evaluate our density estimation methods using $\log p(x)$ as a detection measure. In the case of text, $\log p(x)$ is defined as $\sum_{i=1}^{t} \log p(x_i \mid x_{<i})$.

While *PPL* accounts for varying sequence lengths by averaging word likelihoods over the input sequence, $\log p(x)$ does not. Figure 4 shows that this difference significantly impacts performance. With IMDB as the ID data, using $\log p(x)$ fails for SST-2, achieving close to 100 FAR95 and near 0 AUROC. We suspect this because IMDB examples are a full paragraph while SST-2 examples are one to two sentences. $\log p(x)$ would naturally be smaller for IMDB examples than these OOD examples, resulting in complete failure for simple thresholding methods measured by AUROC.

B FAR95 Results

We additionally evaluate the performance for all experiments using FAR95, which measures the false positive rate at 95% recall. In the context of OOD detection, this measure gives the misclassification rate of ID data at 95% recall of OOD classification, hence a lower value indicates better performance.

Tables 5, 6, 7 and 8 show the results obtained using FAR95 as a metric for the corresponding ID/OOD pairs used earlier. We observe that FAR95 results are in line with AUROC results except for

ID	OOD	FAR95 (↓) *PPL*	*MSP*	Oracle
Fiction	Government	**57.4**	95.0	9.7
	Slate	**66.0**	92.7	37.7
	Telephone	**29.1**	93.3	36.0
	Travel	**58.0**	93.3	10.0
Government	Fiction	**74.7**	92.6	6.4
	Slate	**70.7**	92.1	13.7
	Telephone	**35.2**	95.5	6.2
	Travel	**52.8**	92.4	6.2
Slate	Fiction	**90.6**	96.2	32.2
	Government	**90.0**	96.1	12.6
	Telephone	**57.4**	96.0	22.7
	Travel	**83.3**	95.8	16.8
Telephone	Fiction	**54.2**	93.3	32.5
	Government	**50.9**	93.7	8.5
	Slate	**49.6**	91.1	36.3
	Travel	**44.6**	91.4	10.7
Travel	Fiction	**74.5**	95.5	10.2
	Government	**69.0**	94.4	7.8
	Slate	**75.9**	93.8	16.8
	Telephone	**30.3**	93.7	9.5

Table 7: FAR95 scores obtained using *PPL*, *MSP* and Oracle for background shift caused by shift in MNLI genre. For each pair, lower score obtained (by *PPL* or *MSP*) is in **bold**.

ID	OOD	Shift	FAR95 (↓) *PPL*	*MSP*	Oracle
IMDB	c-IMDB	Semantic	93.1	**82.8**	**69.3**
MNLI	HANS	Background	**4.2**	73.1	0.0
	Negation	Background	94.9	**93.5**	0.1
	Len. Mismatch	Background	98.3	**95.0**	0.1
	Spell. Error	Background	96.9	**92.4**	3.0
	Word Overlap	Background	96.0	**94.4**	1.1
	Antonym	Semantic	100.0	**90.8**	6.3
	Num. Reas.	Semantic	99.5	**77.6**	0.7

Table 8: FAR95 scores obtained using *PPL*, *MSP* and Oracle for challenge data. The primary type of shift observed is indicated in the 'Shift' column. Lower score (among *MSP*/*PPL*) for each pair is in **bold**.

DBPedia, in which case density estimation methods yield a better result. The difference may be a result of the accumulative nature of AUROC in contrast to FAR95, which is a point measurement.

C Background shift in MNLI Genres

MNLI is a crowd-sourced collection of sentence pairs for textual entailment sourced from 10 genres including Fiction, Government, Slate, Telephone, and Travel. We use examples from these five MNLI genres and separately consider each genre as ID and OOD, using the validation splits for evaluation.

Table 9 shows the results for MNLI genres. The discriminative model generalizes well to other genres and we find that the OOD detection performance of calibration method is close to random (50) because of the higher confidence on correct OOD predictions by a well-calibrated model.

ID	OOD	AUROC *PPL*	*MSP*	Oracle	Accuracy OOD	ID
Fiction	Govt.	**83.3**	48.5	98.4	87.0	86.1
	Slate	**81.6**	54.1	92.7	82.2	
	Tel.	**92.3**	51.0	94.6	84.0	
	Travel	**82.2**	49.9	98.3	84.3	
Govt.	Fiction	**75.2**	57.4	98.9	82.8	88.4
	Slate	**77.1**	58.3	97.7	82.0	
	Tel.	**89.9**	57.6	98.5	82.8	
	Travel	**82.6**	57.1	99.4	84.1	
Slate	Fiction	**60.6**	48.2	94.2	84.3	82.5
	Govt.	**61.3**	45.3	97.7	87.8	
	Tel.	**83.0**	49.8	95.3	84.1	
	Travel	**63.7**	46.8	97.6	84.6	
Tel.	Fiction	**85.7**	55.9	95.2	82.5	85.7
	Govt.	**86.0**	52.5	98.3	85.9	
	Slate	**86.8**	59.2	94.2	80.6	
	Travel	**87.8**	56.8	98.6	82.5	
Travel	Fiction	**76.4**	54.8	98.0	81.3	86.7
	Govt.	**78.8**	49.0	98.7	87.4	
	Slate	**77.2**	55.8	96.3	80.8	
	Tel.	**92.7**	56.0	98.1	82.2	

Table 9: Performance on background shifts caused by shift in MNLI genre. For each pair, higher score obtained (by *PPL* or *MSP*) is in **bold**. We can see that the density estimation method using *PPL* significantly outperforms the calibration method.