BARLE: Background-Aware Representation Learning for Background Shift Out-of-Distribution Detection

Hanyu Duan,[1] **Yi Yang,**[1] **Ahmed Abbasi,**[2,3] **Kar Yan Tam**[1]
[1] Department of Information Systems, Business Statistics and Operations Management, HKUST
[2] Human-centered Analytics Lab, University of Notre Dame
[3] Department of IT, Analytics, and Operations, University of Notre Dame
hduanac@connect.ust.hk, {imyiyang,kytam}@ust.hk, aabbasi@nd.edu

Abstract

Machine learning models often suffer from a performance drop when they are applied to out-of-distribution (OOD) samples, i.e., those drawn far away from the training data distribution. Existing OOD detection work mostly focuses on identifying semantic-shift OOD samples, e.g., instances from unseen new classes. However, background-shift OOD detection, which identifies samples with domain or style-change, represents a more practical yet challenging task. In this paper, we propose **B**ackground-**A**ware **R**epresentation **Le**arning (BARLE) for background-shift OOD detection in NLP. Specifically, we generate semantics-preserving background-shifted pseudo OOD samples from pretrained masked language models. We then contrast the in-distribution (ID) samples with their pseudo OOD counterparts. Unlike prior semantic-shift OOD detection work that often leverages an external text corpus, BARLE only uses ID data, which is more flexible and cost-efficient. In experiments across several text classification tasks, we demonstrate that BARLE is capable of improving background-shift OOD detection performance while maintaining ID classification accuracy. We further investigate the properties of the generated pseudo OOD samples, uncovering the working mechanism of BARLE.

1 Introduction

Most state-of-the-art NLP models are evaluated with the assumption that the training data and testing data is drawn from the same distribution. However, when models are deployed in real-world settings, this assumption can be easily violated, and current NLP models tend to suffer from drastic performance drops on out-of-distribution (OOD) data (Hein et al., 2019; Hendrycks and Gimpel, 2016; Nguyen et al., 2015). Identifying OOD samples and distinguishing them from in-distribution (ID) ones, known as OOD detection, plays an essential role in a wide range of NLP applications (Kamath et al., 2020; Kumar and Sarawagi, 2019; Mukherjee and Awadallah, 2020).

Existing OOD detection methods in NLP mostly focus on identifying semantic-shift OOD samples (e.g., samples from unseen classes) (Yilmaz and Toraman, 2020; Zhan et al., 2021; Shu et al., 2017). However, Arora et al. (2021) point out that it is rare to encounter semantic-shift OOD inputs in real-world settings. In practice, background-shift OODs may be more pervasive. These are samples that belong to the same task as the ID data, but with a shift on background features, such as changes in the domain or the style of the text (Ren et al., 2019). For instance, a topic classification model trained on news articles (ID) vs. tweet messages (OOD), or a sentiment classification model trained on movie reviews (ID) vs. product reviews (OOD). Background-shift OOD detection represents a more practical yet challenging task. For example, consumer-facing manufacturers are routinely interested in building sentiment classification models to understand consumer sentiment for product-related issues in online reviews. However, such reviews also contain non-product design/aspect related reviews (such as retailer shipping and customer service experiences). If the sentiment classification model is applied to this review dataset without identifying the background-shift OODs (i.e., non-product design/aspect related reviews), the model may result in a lower prediction performance and thus negatively affect the company's decision on product-related issues. Other applications of background-shift OOD include psychometric NLP tasks such as inferring users' trust, anxiety, and literacy across domains like health and finance (Ahmad et al., 2020; Abbasi et al., 2021).

In this work, we propose an efficient and effective approach for background-shift OOD detection. First, our approach does not require any external data. Prior OOD detection methods often rely on external text corpora to simulate the OOD sam-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 750–764
December 7-11, 2022

ples and learn ID-specific representations (Chen and Yu, 2021; Hendrycks et al., 2018; Xu et al., 2021). However, it is difficult to decide which external data to use, and the choice of the external dataset is critical for successful OOD detection (Hendrycks et al., 2018). Moreover, incorporating prior knowledge in choosing OOD data for training may introduce inductive bias to the OOD detector (Arora et al., 2021). In our work, we utilize pretrained language models (e.g., DistilBERT (Sanh et al., 2019), BERT (Devlin et al., 2018)) and perform a masked language modeling heuristic to generate semantics-preserving background-shifted pseudo samples from the ID data. For instance, a sentiment classification ID example "a sane and breathtakingly creative film" might be mapped to a pseudo example "a massive and perfectly executed painting", where the background features (*film*) are replaced but the positive semantics is preserved. By taking advantage of pretrained language models, we can obtain better quality pseudo OODs in a more principled manner.

Second, we design a background-aware contrastive loss to push the ID training samples apart from their pseudo OOD counterparts. Combined with another semantic contrastive loss, the learned representations are not only semantically distinguishable (which is important for the main task) but also encode rich ID background information (which is important for OOD detection). We then employ an existing OOD scoring mechanism (Hendrycks and Gimpel, 2016; Liu et al., 2020; Lee et al., 2018; Zhou et al., 2021) to map the learned background-aware representation to a scalar that indicates the OOD likelihood.

Our approach is named BARLE, short for **B**ackground-**A**ware **R**epresentation **Le**arning for identifying background-shift OODs. In experiments across several text classification tasks, we show that BARLE achieves superior performance in identifying background-shift OOD samples while maintaining the ID task performance. This implies that BARLE can be used as a 2-in-1 model which not only delivers desirable performance for the ID task, it can also detect any suspicious OOD samples. We also investigate the properties of generated pseudo OOD samples to better understand the working mechanism of BARLE, which may shed light on future OOD detection work. The code is publicly available via GitHub. [1]

[1] https://github.com/hduanac/BARLE. Our implementation is adapted from Zhou et al. (2021), we greatly appreciate the authors for releasing the code to the community.

2 Related Work

Out-of-Distribution Detection. Out-of-distribution (OOD) detection is one of the essential ingredients in building safe and reliable intelligent systems (Amodei et al., 2016; Caruana et al., 2015; Eykholt et al., 2018). OOD detection aims at identifying examples that diverge from the training distribution during inference. According to the literature (Ren et al., 2019; Yang et al., 2021; Arora et al., 2021), OOD samples can be categorized into semantic-shift OODs and background-shift OODs, based on whether the distribution shift is dominated by changes in the semantic or background features, respectively. Semantic-shift OOD detection assumes that the class of an OOD sample does not belong to any of the ID classes. Thus, some OOD detection work directly uses the model's confidence score output (such as softmax probabilities) to indicate the likelihood of OODs (Hendrycks and Gimpel, 2016; Liu et al., 2020; Granese et al., 2021). Instead of directly leveraging the model's outputs, other approaches learn a predictive model to detect OOD samples (Li et al., 2021; Zhan et al., 2021; Chen and Yu, 2021; Xu et al., 2020; Chen et al., 2020a; Yan et al., 2020; Mohseni et al., 2020). Compared to the semantic-shift OODs, samples with background shifts are more common and are also more difficult to identify (Arora et al., 2021). However, despite its prevalence, background-shift OOD detection has attracted limited attention. Our work aims to fill this research gap.

Contrastive Learning for OOD Detection. Contrastive learning has shown remarkable success in representation learning across different domains and tasks (Chen et al., 2020b; He et al., 2020; Giorgi et al., 2020), and has also been used for OOD detection in NLP (Zeng et al., 2021a,b; Zhou et al., 2022, 2021; Mou et al., 2022). However, most prior works using contrastive learning focus on semantic-shift OOD detection, where the contrastive OOD samples are usually constructed from external corpora belonging to different NLP tasks. Our work is novel in that we study background-shift OOD detection - the contrastive samples and contrastive learning objectives are different from prior work.

Data Generation with PLMs. Recent research has shown great interest in leveraging pretrained

language models (PLMs) to generate data for enhancing the performance of NLP tasks (Min et al., 2021), such as information extraction (IE) (Guo and Roth, 2021; Veyseh et al., 2021), sentiment analysis (SA) (Yu et al., 2021; Li et al., 2020), dataset generation (Schick and Schütze, 2021) and few shot learning (Schick and Schütze, 2020). To the best of our knowledge, no prior work has leveraged PLMs to augment ID samples and facilitate OOD detection, and our work demonstrates that this is a viable solution.

3 Method

3.1 Problem Formulation

We now formally define the background-shift OOD detection task. For a given dataset $\mathcal{D}_{ID} = \{(\boldsymbol{x}_m, y_m)\}_{m=1}^{M}$ sampled from a data distribution $P_{ID}(\mathcal{X}, \mathcal{Y})$ (i.e., in-distribution), our goal is to build an OOD detector from $\mathcal{D}_{ID}$ to identify whether an arbitrary input $\boldsymbol{x}$ is drawn from the ID data distribution (i.e., $P_{ID}(\mathcal{X}, \mathcal{Y})$) at inference time or not. We consider an input $(\boldsymbol{x}, y)$ to be a background-shift OOD sample if $(\boldsymbol{x}, y) \sim P_{OOD}(\mathcal{X}, \mathcal{Y}) \neq P_{ID}(\mathcal{X}, \mathcal{Y})$, i.e., it is generated from a data distribution other than the ID data distribution $P_{ID}(\mathcal{X}, \mathcal{Y})$ but its class y belongs to one of the ID classes. We aim to learn an encoder $\phi(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathcal{H}$ that maps an instance $\boldsymbol{x}$ to a hidden representation $\boldsymbol{h} \in \mathcal{H}$. Then an OOD scoring mechanism further maps the hidden representation $\boldsymbol{h}$ to a scalar indicating the likelihood of the input $\boldsymbol{x}$ being OOD.

3.2 An Overview of BARLE

The overall framework of our proposed BARLE method is shown in Figure 1, and the procedure pseudocode is presented in Algorithm 1. It is composed of two major phases. In the first phase (§3.3), we generate pseudo OOD samples using ID training data by performing masked language modeling (MLM) with pretrained language models. This phase aims at synthesizing semantics-preserving background-shifted pseudo OOD samples. In the second phase (§3.4), we use contrastive learning on both the pseudo OOD samples and the ID training samples. We hope to learn background-aware and semantic-aware representations that can benefit not only the OOD detection but also the main task. Finally, given the learned representations, an OOD scoring mechanism is applied to identify the OOD likelihood.

Algorithm 1 Background-Aware Representation Learning (BARLE)

Input: ID training set $\mathcal{D}_{ID}$ and ID dev set $\mathcal{D}_{dev}$.
Output: Main task classifier with OOD detector.

```
/* Initialization step*/
Load the generator G and main task model M.
/* Pseudo OOD generation step */
for x in D_ID do
    Generate x^pseudo for x using G.
    Add x^pseudo to D_pseudo.
/* Contrastive representation learning step */
for t = 1, ... T do
    Sample from D_ID and D_pseudo ∪ D_ID.
    Calculate L_CE, L_conS, and L_conB.
    Total loss: L = L_CE + αL_conS + βL_conB.
    Update the model parameters by L.
    /* Evaluation step */
    if t%steps_eval == 0 do
        Evaluate M with D_dev.
Return the best model.
```

3.3 Pseudo OOD Sample Generation

We leverage a pretrained masked language model to generate pseudo OOD samples using ID training data in a principled manner. To facilitate background-shift OOD detection, we expect to generate semantics-preserving background-shifted pseudo samples. Specifically, for a given instance $\boldsymbol{x}$ from the ID training set, we take a pretrained masked language model (such as BERT or DistilBERT), denoted as the generator G, to produce a corresponding pseudo sample $\boldsymbol{x}^{pseudo}$.

The generation process works as follows. The first step is to perform token masking. Given an instance $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ as the input to the generator G, say a sentence with n tokens, we randomly select one position m (an integer between 1 and n) to mask out. The token of the selected position m is replaced with a `[MASK]` token. We denote the masked instance as $\boldsymbol{x}^{masked} = REPLACE(\boldsymbol{x}, m, \texttt{[MASK]})$. The second step is to predict the masked token. The generator produces an output distribution over all the tokens in the vocabulary for that masked-out position, i.e., $P_G(x_m \mid \boldsymbol{x}^{masked})$. We sample a token from this distribution (i.e., $\hat{x}_m \sim P_G(x_m \mid \boldsymbol{x}^{masked})$) to replace the original token, i.e., $\boldsymbol{x}^{pseudo} = REPLACE(\boldsymbol{x}, m, \hat{x}_m)$. Instead of sampling from the entire vocabulary, we sample the target token from a candidate set composed

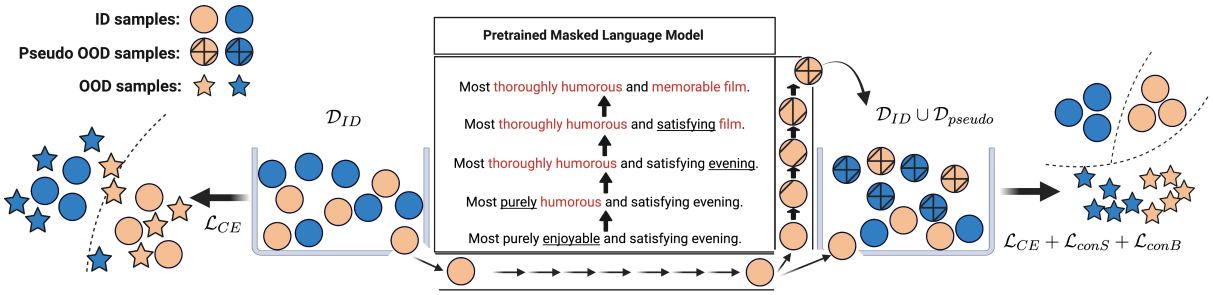

Figure 1: Illustration of BARLE process. Pseudo OODs are generated from pretrained masked language models, and contrastive losses are applied to learn background-aware representations. The traditional cross-entropy training scheme may fail to detect background-shift OODs.

of the tokens with top-k highest probabilities because we want to avoid syntactic and semantic errors in the generated text. We apply the two steps iteratively until the replacement ratio ρ, the percentage of replaced tokens, achieves a pre-defined threshold. Finally, we add both the ID examples and the corresponding OOD samples together as $\mathcal{D}_{ID} \cup \mathcal{D}_{pseudo}$ for subsequent use.

We provide some examples of the generated pseudo OOD samples in Table 1. The examples provide us an intuitive understanding of the proposed pseudo OOD generation mechanism. We can observe that the background features such as *movie* and *film* of the ID data shift to diverse domain features such as *painting* and *web*, whereas the sentiment features are well-preserved in the generated pseudo OOD samples. Moreover, we also empirically show, in the experiment section, that the generated pseudo samples data can indeed preserve semantics but with background shift (See Figure 2).

3.4 Contrastive Representation Learning

We now present how to effectively utilize the pseudo OOD samples (§3.3) for background-shift OOD detection.

Contrasting Background-Shifted Instances. In prior works, most contrastive learning schemes applied to OOD detection problems act by pulling instances from the same semantic class closer while pushing samples with different class labels apart (Zeng et al., 2021a,b; Zhou et al., 2021). However, this may fail to detect background-shift OOD samples because such learning processes are mainly based on contrasting semantic features, while the background features are ignored. In other words, prior semantic-shift OOD detection works may very well distinguish a sentiment review from a machine translation text, but they may not tell a movie review apart from a restaurant review.

To address this issue, we propose to contrast the ID training samples with their semantics-preserving background-shifted augmentations (i.e., the pseudo OOD samples) to encode the ID background information into the learned representations. We utilize a margin-based contrastive loss (Chopra et al., 2005). Specifically, we first sample a batch of instances $\mathcal{B} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ from $\mathcal{D}_{ID} \cup \mathcal{D}_{pseudo}$, and let $\boldsymbol{x}_i$ be a query instance drawn from the batch. If $\boldsymbol{x}_i$ is an ID sample, all the ID samples in the batch except the query instance construct the positive set $\{\boldsymbol{x}_i^+\}$, and the negative set $\{\boldsymbol{x}_i^-\}$ is composed of all the pseudo OOD samples in the batch. Similarly, if $\boldsymbol{x}_i$ is a pseudo OOD sample, the positive set consists of all the pseudo OOD samples in the batch except $\boldsymbol{x}_i$, and all the ID samples in the batch construct the negative set. Formally, we denote the anchor set of the query instance $\boldsymbol{x}_i$ as $\mathcal{A}(\boldsymbol{x}_i) = \mathcal{B} \setminus \boldsymbol{x}_i$. Then the positive set and the negative set are defined as $\{\boldsymbol{x}_i^+\} = \{\boldsymbol{p} \in \mathcal{A}(\boldsymbol{x}_i) : y_i = y_p\}$ and $\{\boldsymbol{x}_i^-\} = \{\boldsymbol{n} \in \mathcal{A}(\boldsymbol{x}_i) : y_i \neq y_n\}$, respectively. The background contrastive loss is formulated as:

$$\mathcal{L}_{pos}(\boldsymbol{x}_i, \{\boldsymbol{x}_i^+\}) = \sum_{\boldsymbol{x}_i' \in \{\boldsymbol{x}_i^+\}} \| \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i') \|^2, \quad (1)$$

$$\mathcal{L}_{neg}(\boldsymbol{x}_i, \{\boldsymbol{x}_i^-\}) = \sum_{\boldsymbol{x}_i' \in \{\boldsymbol{x}_i^-\}} (\xi - \| \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i') \|^2)_+, \quad (2)$$

$$\mathcal{L}_{conB} = \frac{1}{N}\Big(\sum_{\boldsymbol{x} \in \mathcal{B}} \frac{1}{|\{\boldsymbol{x}^+\}|} \mathcal{L}_{pos}(\boldsymbol{x}, \{\boldsymbol{x}^+\}) + \sum_{\boldsymbol{x} \in \mathcal{B}} \frac{1}{|\{\boldsymbol{x}^-\}|} \mathcal{L}_{neg}(\boldsymbol{x}, \{\boldsymbol{x}^-\})\Big), \quad (3)$$

where the margin ξ is defined following the prior work (Zhou et al., 2021) as the maximum distance between pairs of instances from the same class in a batch, and $\phi(\boldsymbol{x})$ denotes the hidden representation (i.e., the input to the softmax layer) of an instance.

Original ID examples	Generated pseudo OOD examples
the great films about movie love. (+)	the great songs about philippine theater. (+)
the film runs on a little longer than it needs to. (-)	the lawsuit stalled on a price cheaper than it intended to. (-)
the movie is brilliant, really. (+)	the decor is clean, flawless. (+)
without having much dramatic impact. (-)	without having real economic implications. (-)
an extremely unpleasant film. (-)	an extremely sour odor. (-)
a sane and breathtakingly creative film. (+)	a massive and perfectly executed painting. (+)
this disappointed by a movie in a long time. (-)	this follows by a break in a limited budget. (-)
it 's still unusually crafty and intelligent for hollywood horror. (+)	it's therefore fairly creative and exciting for web designers. (+)

Table 1: Generated pseudo OOD examples with their corresponding ID examples sampled from the SST2 dataset. The "+" denotes the positive sentiment label and "-" denotes the negative sentiment label. We highlight (underline) the potential background words with blue (dashed line) and sentiment words with red (solid line).

Contrasting Semantically Different Instances. We use another contrastive loss that contrasts semantically different instances. The idea is to learn compact semantic representation clusters, which may facilitate the main task performance. Similarly, we define the semantics contrastive loss as:

$$\mathcal{L}_{conS} = \frac{1}{N}\Big(\sum_{\boldsymbol{x}\in\mathcal{B}_{ID}} \frac{1}{|\{\boldsymbol{x}^+\}|}\mathcal{L}_{pos}(\boldsymbol{x},\{\boldsymbol{x}^+\}) + \sum_{\boldsymbol{x}\in\mathcal{B}_{ID}} \frac{1}{|\{\boldsymbol{x}^-\}|}\mathcal{L}_{neg}(\boldsymbol{x},\{\boldsymbol{x}^-\})\Big), \quad (4)$$

where $\mathcal{B}_{ID}$ denotes a batch of instances sampled from the ID training set $\mathcal{D}_{ID}$, and the positive set and negative set are constructed based on the class label of the main task. The model is also trained with cross-entropy loss $\mathcal{L}_{CE}$. We formulate the overall total training loss as:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{conS} + \beta\mathcal{L}_{conB}, \quad (5)$$

where α and β control the strengths of the semantics contrasting and the background contrasting, respectively, both of which are tuned on the ID development set. With this loss, we expect the model to learn background-aware and task-specific representations that benefit not only the OOD detection but also the main task.

Finally, we use an OOD scoring mechanism to map the learned hidden representation to a scalar indicating the likelihood of the instance being an OOD sample. Various scoring mechanisms have been proposed such as MSP (Hendrycks and Gimpel, 2016), Energy (Liu et al., 2020), MHLNB (Lee et al., 2018) and Cosine (Zhou et al., 2021), and our proposed training scheme is scoring mechanism-agnostic. A brief introduction of these four scoring mechanisms and how they work with BARLE are presented in Appendix A.

4 Experiments

4.1 Datasets

We consider two NLP tasks in the experiments: topic categorization and sentiment classification. For topic categorization, we use the **Yahoo-AGNews-five** testbed (Li et al., 2021). This dataset is composed of a subset of Yahoo!Answers as the ID data, and a subset of AGNews Corpus as the OOD data [2]. The two datasets share the same label space, but their text style shifts.

For sentiment classification, we use three popular datasets (i.e., SST2, IMDB, and Amazon). Among the three datasets, **SST2** and **IMDB** encompass movie reviews, whereas **Amazon** includes online consumer reviews of Amazon products (Blitzer et al., 2007). For the Amazon data, consistent with prior work, we retain the reviews from four categories (i.e., Books, DVDs, Electronics, and Kitchen appliances) to simulate the text domain shift scenario. The statistics of these datasets are shown in Table 2. These datasets have all been used in prior work for benchmarking semantic shift OOD detection (Hendrycks et al., 2020; Li et al., 2021).

Dataset	# ID training	# ID dev	# ID test	# OOD test	# Class
Yahoo-AGNews-five	10,000	2,500	2,500	2,500	5
SST2	67,349	872	1821	-	2
IMDB	22,500	2,500	25,000	-	2
Amazon	6,400	800	800	-	2

Table 2: Statistics of the datasets.

We then construct the ID/OOD dataset pairs for background-shift OOD detection by pairing the datasets belonging to the same task, summarized in Table 3. We train the model on the ID training set, and the ID development set is used for param-

[2]The ID instances are selected from 5 classes (i.e., "Health", "Science & Mathematics", "Sports", "Entertainment & Music", and "Business & Finance") of the original Yahoo!Answers, and the OOD set is constructed by the samples of the original AG Corpus from "Health", "Sci/Tech", "Sports", "Entertainment", and "Business".

eter tuning. We evaluate the model's main task performance on the ID test set. The OOD detection performance is assessed on the OOD test set.

Task	ID dataset	OOD dataset
Topic categorization	Yahoo!Answers	AGNews
Sentiment classification	SST2	IMDB
		Amazon
	Amazon	IMDB
		SST2
	Amazon-Books	Amazon-DVDs
		Amazon-Electronics
		Amazon-Kitchen

Table 3: ID/OOD setups in the experiments.

4.2 Evaluation Metrics and Benchmarks

Evaluation Metrics. For OOD detection, we consider two commonly used metrics following prior work (Hendrycks and Gimpel, 2016; Lee et al., 2018), i.e., the AUROC and the FAR95. **AUROC** measures how much the model can distinguish the OOD samples from ID samples. Higher AUROC scores indicate better OOD detection capabilities, and a random guess detector would have an AUROC score of 0.5. **FAR95** can be interpreted as the probability that an OOD sample (negative) is misclassified by the detector as an ID sample (positive) when the true positive rate (TPR) is equal to 95%. A lower FAR95 value indicates better OOD detection performance. We use classification accuracy (**ACC**) as the main task metric.

Benchmarks. Prior work has empirically showed that existing approaches for semantic-shift OOD detection perform poorly on background-shift OOD detection tasks, and there is little work on background-shift OOD detection (Arora et al., 2021; Li et al., 2021). Therefore, to choose suitable benchmarks to compare with, we consider the following two approaches. We first consider training the ID model using the vanilla cross-entropy loss on the ID data, only. We denote this benchmark as "Vanilla" (**VAN**). The second benchmark follows existing semantic shift OOD detection work by using an arbitrarily chosen external dataset as the pseudo OODs. We denote this benchmark as "External" (**EXT**). Following (Zhou et al., 2021), we choose the English text of a machine translation dataset (i.e., English-German WMT16 (Bojar et al., 2016)) as the auxiliary external data for EXT. In addition, we consider a density-based method (**PPL**) following (Arora et al., 2021). Specifically, we fine-tune GPT-2 (Radford et al., 2019) on the original ID data, and use the token perplexity as the OOD score. We implement the officially pretrained GPT-2_{Small} using the `transformers` library [3].

4.3 Hyperparameter Settings

For pseudo OOD generation, we use the replacement ratio in the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and set the candidate size to 100. We do not substitute stopwords and synthesize pseudo OOD samples for all the ID training instances. For the representation learning, we build the classifier upon the officially pretrained RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2018) with different model scales using the `transformers` library[4]. The model is optimized by AdamW (Loshchilov and Hutter, 2017) with 0.01 weight decay and 0.06 warmup ratio. We choose the learning rate from the range of $[1e-7, 1e-6]$, and we use a batch size of 8. The maximum sequence length is set to 256, and the parameters are tuned based on the contrastive loss and the classification performance on the ID development set.

To make it easier for practitioners to integrate BARLE in their working pipelines, we provide a guidance on hyperparameter selection as follows. For the replacement ratio ρ, we recommend adjusting it based on the average length of input text. In general, we recommend setting a larger ratio (e.g., 0.5) for longer input text (e.g., hundreds of words on average) and smaller ratio (e.g., 0.1) for those with only dozens of words. For the number of generated pseudo OOD samples, we generate one pseudo OOD for one ID training example. As noted (§6), selecting informative pseudo OODs for efficient OOD detection may constitute an interesting future direction. For the top candidate k, as we demonstrate in Figure 3, values that are too large or too small will not generate favourable pseudo OOD examples for the subsequent detection. We recommend a moderate size of 100 as a sensible default value.

4.4 Background-Shift OOD Detection Results

We use DistilBERT (Sanh et al., 2019) as the generator to synthesize pseudo OOD samples [5] using the `nlpaug` library (Ma, 2019). We tune and set the contrastive loss hyperparameter α and β to

[3] https://huggingface.co/gpt2

[4] https://github.com/huggingface/transformers

[5] We also use BERT (Devlin et al., 2018) as the generator and find that the results are consistent. Thus we use DistilBERT in the experiments for its small scale and the efficiency purpose.

1.5 and 2.0 in the experiments. We use four different OOD scoring mechanisms, including MSP (Hendrycks and Gimpel, 2016), Energy (Liu et al., 2020), MHLNB (Lee et al., 2018) and Cosine (Zhou et al., 2021). Details about the scoring mechanisms appear in Appendix A.

The main experimental results (with RoBERTa-Large as the underlying model) appear in Table 4, Table 5 and Table 6. Full results using other pretrained models (RoBERTa-Base and BERT-Large/Base) are presented in Appendix B. First, we see that BARLE significantly outperforms benchmarks VAN and EXT on OOD detection, and the performance lift is consistent across tasks and datasets. The improvement over EXT indicates that our pseudo OOD generation is more effective than an arbitrary external dataset. Second, we find that BARLE gets more significant gains when combined with the density-based scoring mechanisms (i.e., MHLNB, and Cosine) compared with the calibration-based ones (MSP and Energy). We interpret this as follows: both density-based methods and our proposed contrastive losses work on the same hidden representation space so that such scoring mechanisms can directly benefit from the contrastively learned representations. This is consistent with (Arora et al., 2021) that density estimation methods can better account for background information shifts. We also examine the model's main task classification accuracy on the ID testbed. The results appearing in Appendix C show that BARLE maintains desirable main task performance.

ID dataset	Model	OOD metrics	
		AUROC ↑	FAR95 ↓
	AGNews		
	PPL		
	GPT-2	0.596	0.942
	VAN		
	RoBERTa-Large w/ MSP	0.696	0.864
	RoBERTa-Large w/ Energy	0.740	0.778
	RoBERTa-Large w/ MHLNB	0.888	0.436
Yahoo!Answers	RoBERTa-Large w/ Cosine	0.796	0.630
	EXT		
	RoBERTa-Large w/ MSP	0.704	0.885
	RoBERTa-Large w/ Energy	0.775	**0.722**
	RoBERTa-Large w/ MHLNB	0.945	0.288
	RoBERTa-Large w/ Cosine	0.883	0.498
	BARLE		
	RoBERTa-Large w/ MSP	**0.747**	**0.853**
	RoBERTa-Large w/ Energy	**0.788**	0.764
	RoBERTa-Large w/ MHLNB	**0.960**	**0.218**
	RoBERTa-Large w/ Cosine	**0.922**	**0.337**

Table 4: Topic categorization OOD detection.

Ablation Study To conduct an ablation analysis on the contrastive objectives, we turn off the semantics loss and background loss by setting α and β in Equation 5 to zero, respectively. The results on topic categorization and the SST2/IMDB pair appear in Table 7. From the results, we can see that applying either the semantics-contrastive loss or the background-contrastive loss can outperform the one with the cross-entropy loss applied only. Moreover, applying the background-contrastive loss are more effective than applying the semantics-contrastive loss for background-shift OOD detection, and applying both can further improve the performance. The results on all eight ID/OOD pairs are presented in Appendix D.

ID dataset	Model	OOD metrics			
		AUROC ↑	FAR95 ↓	AUROC ↑	FAR95 ↓
		IMDB		*SST2*	
	PPL				
	GPT-2	0.504	0.890	0.932	0.324
	VAN				
	RoBERTa-Large w/ MSP	0.600	0.934	0.950	0.302
	RoBERTa-Large w/ Energy	0.592	0.931	0.964	0.216
	RoBERTa-Large w/ MHLNB	0.537	0.964	0.947	0.411
Amazon	RoBERTa-Large w/ Cosine	0.545	0.964	0.837	0.666
	EXT				
	RoBERTa-Large w/ MSP	0.596	**0.926**	0.971	0.085
	RoBERTa-Large w/ Energy	0.568	0.932	0.962	**0.210**
	RoBERTa-Large w/ MHLNB	0.727	0.904	0.982	0.059
	RoBERTa-Large w/ Cosine	0.653	0.932	0.971	0.140
	BARLE				
	RoBERTa-Large w/ MSP	**0.618**	0.931	**0.978**	**0.034**
	RoBERTa-Large w/ Energy	**0.604**	**0.931**	**0.968**	0.236
	RoBERTa-Large w/ MHLNB	**0.803**	**0.650**	**0.992**	**0.003**
	RoBERTa-Large w/ Cosine	**0.722**	**0.813**	**0.985**	**0.058**
		IMDB		*Amazon*	
	PPL				
	GPT-2	0.919	0.271	0.875	0.486
	VAN				
	RoBERTa-Large w/ MSP	0.756	0.890	0.744	0.926
	RoBERTa-Large w/ Energy	0.770	**0.886**	0.745	**0.891**
	RoBERTa-Large w/ MHLNB	0.993	0.018	0.989	0.043
SST2	RoBERTa-Large w/ Cosine	0.989	0.043	0.982	0.089
	EXT				
	RoBERTa-Large w/ MSP	0.908	0.600	0.769	0.793
	RoBERTa-Large w/ Energy	0.861	0.997	0.740	0.968
	RoBERTa-Large w/ MHLNB	0.999	0.001	0.980	0.064
	RoBERTa-Large w/ Cosine	0.997	0.006	0.909	0.238
	BARLE				
	RoBERTa-Large w/ MSP	**0.927**	**0.520**	**0.845**	**0.706**
	RoBERTa-Large w/ Energy	**0.870**	0.996	**0.795**	0.975
	RoBERTa-Large w/ MHLNB	**0.999**	**0.001**	**0.998**	**0.010**
	RoBERTa-Large w/ Cosine	**0.999**	**0.003**	**0.994**	**0.020**

Table 5: Sentiment classification OOD detection.

4.5 Analysis of the Pseudo OOD samples

Visualization of feature distributions. Pseudo OOD sample generation using PLMs is a critical component in our work. Here, we compare the pseudo OOD feature distributions with that of the ID data. We use the SST2 dataset in this analysis. Following (Ren et al., 2019), we investigate the semantic feature and background feature distributions respectively. To capture the semantic features, we utilize the VADER lexicon [6] to obtain sentiment scores of ID samples and pseudo OOD samples. We compare their sentiment polarity distributions in Figure 2 (left). We can observe that these two distributions are very well-overlapped, which indicates that the semantic meanings of the ID samples are well-preserved in the generated

[6] https://www.nltk.org/api/nltk.sentiment.vader.html

ID dataset	Model	OOD metrics					
		AUROC ↑	FAR95 ↓	AUROC ↑	FAR95 ↓	AUROC ↑	FAR95 ↓
		Kitchen		*Electronics*		*DVDs*	
		PPL					
	GPT-2	0.524	0.900	0.555	0.850	0.556	0.861
		VAN					
	RoBERTa-Large w/ MSP	0.699	**0.928**	0.724	0.924	0.573	0.947
	RoBERTa-Large w/ Energy	0.647	**0.944**	0.717	**0.913**	0.547	0.948
	RoBERTa-Large w/ MHLNB	0.750	**0.889**	0.786	**0.840**	0.571	0.961
Books	RoBERTa-Large w/ Cosine	0.684	0.928	0.746	0.904	0.503	0.965
		EXT					
	RoBERTa-Large w/ MSP	0.696	0.952	0.727	0.937	0.559	0.939
	RoBERTa-Large w/ Energy	0.699	0.955	0.725	0.932	0.561	0.922
	RoBERTa-Large w/ MHLNB	0.756	0.928	0.779	0.923	0.618	0.940
	RoBERTa-Large w/ Cosine	0.724	0.920	0.763	0.907	0.600	0.923
		BARLE					
	RoBERTa-Large w/ MSP	**0.723**	0.950	**0.770**	**0.923**	**0.575**	**0.919**
	RoBERTa-Large w/ Energy	**0.704**	0.954	**0.762**	0.932	**0.579**	**0.925**
	RoBERTa-Large w/ MHLNB	**0.772**	0.898	**0.804**	0.866	**0.665**	**0.857**
	RoBERTa-Large w/ Cosine	**0.757**	**0.907**	**0.811**	**0.888**	**0.628**	**0.895**

Table 6: Sentiment classification OOD detection, on the subsets (Books, Kitchen, Electronics, and DVDs) of Amazon review data.

Task	α/β	OOD metric		ID metric
		AUROC ↑	FAR95 ↓	ACC ↑
Topic categorization (Yahoo!Answers / AGNews)	- / -	0.780	0.677	0.829
	+ / -	0.792	0.636	0.822
	- / +	0.843	0.553	0.842
	+ / +	**0.854**	**0.543**	0.830
Sentiment classification (SST2 / IMDB)	- / -	0.877	0.459	0.960
	+ / -	0.899	0.438	0.956
	- / +	0.927	0.424	0.947
	+ / +	**0.949**	**0.380**	0.954

Table 7: Ablation study on the contrastive losses. The "-" and "+" denote setting the corresponding parameters to zeros or not respectively. The results are averaged across four scoring mechanisms.

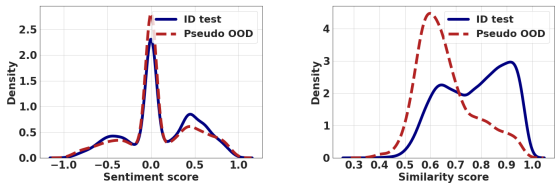

Figure 2: Semantic feature distributions (left) and background similarity distributions (right) for the ID and pseudo OOD samples via kernel density estimation.

pseudo OOD samples. To examine the background features, we first train a doc2vec (Le and Mikolov, 2014) model with the ID training samples for capturing the surface ID background information. For each ID test sample and each pseudo OOD sample, we retrieve their most similar instances in the ID training set and calculate their cosine similarity scores. We plot the similarity score distributions in Figure 2 (right). We can see a clear distributional shift, which reflects the differences of the background statistics between the ID data and the generated pseudo OOD data. This analysis validates that BARLE can generate semantics-preserving background-shifted pseudo OOD samples.

Hard Examples. The quality of pseudo OOD samples is a common concern for OOD detection in prior works, as they often use external text corpus as OODs. In this analysis, we show that our generated pseudo OOD samples are "hard" negative examples whose native representations are located near the ID distribution. This is important because Robinson et al., 2020 show that contrastive learning can benefit from "hard" negative examples. Specifically, we fit a multivariate Gaussian distribution on the ID training (i.e., Amazon) examples' hidden representations. We then measure the distance between the estimated distribution and each instance from the ID test set, the pseudo OOD set, and an arbitrarily chosen external set (i.e., WMT16) using the Mahalanobis distance metric. We visualize the distance distributions on a log scale in Figure 4. It shows that the pseudo OOD samples generated by BARLE are distributed much more closely to the ID samples on the hidden representation space compared with the arbitrarily chosen external instances, which shows that our generated pseudo OOD samples are hard examples. On the contrary, the external data distributions are far away from the ID test, indicating that they are easy negative examples so that contrastive learning may not sufficiently learn ID-specific information. This analysis also reveals that the common strategy of using external text corpora for background-shift OOD detection may not be effective.

Sensitivity Analysis. As part of our sensitivity analysis on pseudo OOD sample generation, we

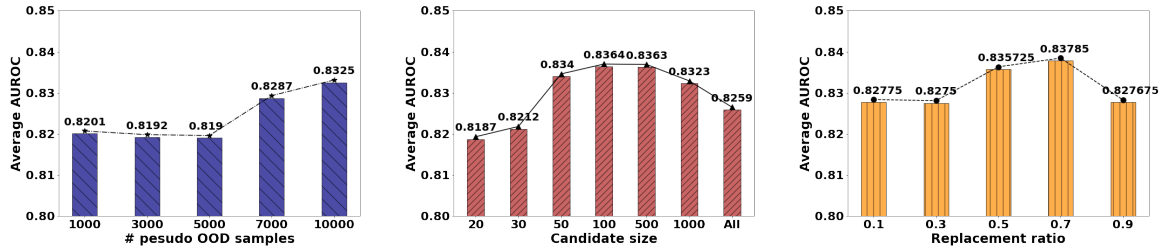

Figure 3: Average OOD detection performances evaluated under different number of pseudo OOD samples (left), candidate sizes (mid), and replacement ratios (right). The candidate size "All" represents the entire vocabulary.

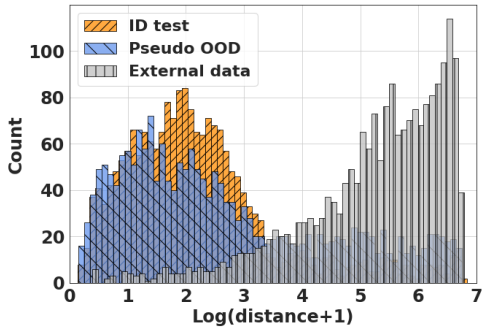

Figure 4: Distance distributions between ID training (Amazon data) and ID test, Pseudo OODs (our approach), and an external dataset respectively.

study the impact of 1). the number of pseudo OOD samples, 2). the candidate size k, and 3). the replacement ratio ρ on the OOD detection performance. The analysis is conducted on the Yahoo-AGNews-five dataset and the results appear in Figure 3. Note that the AUROC score is averaged across the four scoring mechanisms. First, the OOD detection can benefit from a large number of pseudo OOD samples. In the main experiments, we generate one OOD sample for each of the ID samples, which corresponds to 10,000 OOD samples for this task. By decreasing the total number of pseudo OOD samples, we can see that the OOD detection performance decreases (Figure 3 (left)). Second, candidate size k denotes the number of top-k highest probability words to be considered for data augmentation. Figure 3 (mid) shows that neither large candidate size nor small candidate size produces optimal detection performance. This implies that we should adjust the candidate size according to the specific downstream tasks. Third, overly diversified pseudo OOD samples (i.e., large replacement ratio) cannot benefit OOD detection substantially (Figure 3 (right)). This is expected because the semantic features are likely to be changed in the pseudo OOD samples if the replacement ratio is large, and that would violate our objective of having semantics-preserving pseudo samples.

5 Conclusion

In this work, we propose a simple yet effective method for background shift OOD detection. Our method leverages pretrained language models to synthesize semantics-preserving background-shifted pseudo OOD samples from the ID training data. By contrasting the ID training samples with the their pseudo OOD counterparts, our proposed training scheme learns background-aware representations and improves OOD detection performance. Additional analyses on the properties of the generated pseudo OOD samples also validates our design effectiveness. We believe our work sheds new light on OOD detection for building robust NLP systems. As noted, possible applications include an array of NLP-based user modeling tasks including inferring psychometrics, text-based personality detection in forums and social media (Yang et al., 2022), and stylometric authorship identification where style is the primary task and topics and genres are background (Abbasi and Chen, 2008).

6 Limitations

This work has several limitations that can be improved in future research. First, we do not evaluate the effectiveness of our proposed method on semantic-shift OOD detection, since our main focus is background-shift OOD detection. Future work can build upon our pseudo OOD generation approach and further investigate its performance on semantic-shift OOD detection, or develop a unified framework that can detect OODs of different kinds. Second, our approach generates one pseudo OOD for each of the ID training sample. This may not be very efficient for large datasets. Therefore, selecting informative pseudo OODs for efficient OOD

detection may constitute an interesting future direction. Third, our experiments are conducted on two text classification tasks where the label space is generally small (five for topic categorization and two for sentiment classification). How the approach performs in real-world NLP settings where the label space is large, or perhaps even going beyond text classification to prediction/regression tasks, warrants further investigation. Finally, our experiments show that the proposed approach works better with density-based OOD scoring mechanisms than with the calibration-based ones. This is consistent with (Arora et al., 2021) in that density-based methods can better account for background shifts. Future work may be needed to delve deeper into the relationship between OOD representation learning and OOD scoring mechanisms to better understand this OOD phenomenon.

Acknowledgements

We thank anonymous reviewers for their valuable comments and suggestions. This work was funded in part through U.S. NSF grant IIS-2039915 and an Oracle for Research grant entitled "NLP for the Greater Good."

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Ahmed Abbasi, David Dobolyi, John P Lalor, Richard G Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3748–3758.

Faizan Ahmad, Ahmed Abbasi, Jingjing Li, David G Dobolyi, Richard G Netemeyer, Gari D Clifford, and Hsinchun Chen. 2020. A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems (TOIS)*, 38(1):1–29.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Derek Chen and Zhou Yu. 2021. Gold: improving out-of-scope detection in dialogues using data augmentation. *arXiv preprint arXiv:2109.03079*.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2020a. Robust out-of-distribution detection via informative outlier mining. *arXiv preprint arXiv:2006.15207*, 1(2):7.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. 2021. Doctor: A simple method for detecting misclassification errors. *Advances in Neural Information Processing Systems*, 34:5669–5681.

Ruohao Guo and Dan Roth. 2021. Constrained labeled data generation for low-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4519–4533.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. *arXiv preprint arXiv:2004.14769*.

Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. 2021. kFolden: k-fold ensemble for out-of-distribution detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3115, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223.

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. Disentangled knowledge transfer for ood intent discovery with unified contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 46–53.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Uncertainty-aware self-training for text classification with few labels. *arXiv preprint arXiv:2006.15315*.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Advances in Neural Information Processing Systems*, 32.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Timo Schick and Hinrich Schütze. 2020. Few-shot text generation with pattern-exploiting training. *arXiv preprint arXiv:2012.11926*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash gpt-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282.

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. 2020. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. *arXiv preprint arXiv:2106.00948*.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. 2020. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Kai Yang, Raymond YK Lau, and Ahmed Abbasi. 2022. Getting personal: A deep learning artifact for text-based measurement of personality. *Information Systems Research*.

Eyup Halit Yilmaz and Cagri Toraman. 2020. Kloos: Kl divergence-based out-of-scope intent detection in human-to-machine conversations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2105–2108.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021a. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. *arXiv preprint arXiv:2105.14289*.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Hong Xu, and Weiran Xu. 2021b. Adversarial self-supervised learning for out-of-domain detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5631–5639.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. *arXiv preprint arXiv:2106.08616*.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141.

A OOD Scoring Mechanisms

Our proposed method can generate background-aware representations. Given the representations, an OOD scoring mechanism is applied to infer the OOD likelihood. In this appendix, we briefly describe the OOD scoring mechanisms considered in our experiments.

Maximum Softmax Probability (MSP). MSP is first introduced for OOD detection by (Hendrycks and Gimpel, 2016). This method retrieves the maximum class probability from a softmax distribution for calculating the OOD score. Intuitively, higher maximum class probability implies lower likelihood of an OOD. Specifically, the MSP score is defined as:

$$g_{msp} = 1 - \max_{i \in \{1,\dots,C\}} p_i, \tag{6}$$

where C denotes the number of classes in a classification task, and p_i represents the softmax probability for the i-th class. The idea is that a more uniform softmax distribution indicates a higher likelihood to be OOD. Since our method BARLE is also trained with the main task, the output class probability can be directly used by MSP as the OOD likelihood.

Energy Score (Energy). Instead of using the maximum class probability, Liu et al., 2020 propose to measure an energy score of the output probabilities:

$$g_{energy} = -\log \sum_{i=1}^{C} e^{\boldsymbol{w}_i^T \boldsymbol{h}}, \tag{7}$$

where $\boldsymbol{w}_i$ is the weight of the softmax layer in terms of the i-th class, and $\boldsymbol{h}$ denotes the input to the softmax layer, i.e., $\boldsymbol{w}_i^T \boldsymbol{h}$ represents the logit corresponding to the i-th class label. The energy score can better distinguish ID and OOD samples since it is theoretically aligned with the probability density of the inputs, and less sensitive to the over-confidence issue. A higher energy score indicates a higher OOD likelihood. Similar to MSP, the output logits of BARLE are used to calculate the energy score.

Mahalanobis Distance (MHLNB). Lee et al., 2018 fit a class conditional Gaussian distribution on the ID development set $\mathcal{D}_{dev} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ under Gaussian discriminant analysis. They first compute the empirical class mean and covariance by:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \boldsymbol{h}_i, \tag{8}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{c} \sum_{i:y_i=c} (\boldsymbol{h}_i - \hat{\boldsymbol{\mu}}_c)(\boldsymbol{h}_i - \hat{\boldsymbol{\mu}}_c)^T, \tag{9}$$

where c denotes the class label, N_c is the number of instances with the class c, and $\boldsymbol{h}_i$ represents the input to the softmax layer of the i-th instance. Then, the detection score is defined by the Mahalanobis distance between the test sample and the closest class conditional Gaussian distribution, i.e.,

$$g_{MHLNB} = \max_{c} -(\boldsymbol{h} - \hat{\boldsymbol{\mu}}_c)^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{h} - \hat{\boldsymbol{\mu}}_c), \tag{10}$$

where $\boldsymbol{h}$ denotes the hidden representation (i.e., the input to the softmax layer) of the instance. The above metric corresponds to measuring the log of the probability density of the test sample in the estimated Gaussian distribution. The learned background-aware representations from BARLE can be utilized by MHLNB to infer its OOD likelihood.

Cosine Similarity (Cosine). We consider another distance metric based on the cosine similarity (Zhou et al., 2021). Given a test sample, the OOD score is defined by the maximum cosine similarity between the test sample and the instances in the development set:

$$g_{cosine} = -\max_{i \in \{1,\dots,N\}} \cos(\boldsymbol{h}, \boldsymbol{h}_i), \tag{11}$$

where N denotes the size of the development set, and $\boldsymbol{h}$ is the hidden representation, i.e., the input to the softmax layer. Similar to MHLNB, the learned representations from BARLE can be utilized by Cosine to infer its OOD likelihood.

B Additional OOD Detection Results

In the paper, we present the OOD detection results where the NLP tasks are fine tuned on RoBERTa-Large model. In this appendix, we change the underlying model to RoBERTa-Base, BERT-Base, and BERT-Large, and present the OOD results in Table 8 respectively.

C Main Task Performance

Although our focus is OOD detection, we need to examine if the background-aware representation learning will have negative effect on model's main

ID dataset	Model	OOD metrics	
		AUROC ↑	FAR95 ↓
	AGNews		
	VAN		
	RoBERTa-Base w/ MSP	0.707	0.851
	RoBERTa-Base w/ Energy	0.738	0.733
	RoBERTa-Base w/ MHLNB	0.883	0.432
	RoBERTa-Base w/ Cosine	0.796	0.649
	EXT		
	RoBERTa-Base w/ MSP	0.705	0.837
	RoBERTa-Base w/ Energy	0.729	0.714
	RoBERTa-Base w/ MHLNB	0.940	0.322
	RoBERTa-Base w/ Cosine	0.845	0.462
	BARLE		
	RoBERTa-Base w/ MSP	**0.726**	**0.798**
	RoBERTa-Base w/ Energy	**0.756**	**0.694**
	RoBERTa-Base w/ MHLNB	**0.955**	**0.238**
	RoBERTa-Base w/ Cosine	**0.872**	**0.429**
	VAN		
	BERT-Large w/ MSP	0.732	0.867
	BERT-Large w/ Energy	0.695	0.803
	BERT-Large w/ MHLNB	**0.935**	**0.288**
	BERT-Large w/ Cosine	0.792	0.776
	EXT		
Yahoo!Answers	BERT-Large w/ MSP	0.714	0.913
	BERT-Large w/ Energy	0.790	**0.538**
	BERT-Large w/ MHLNB	0.912	0.501
	BERT-Large w/ Cosine	0.779	0.743
	BARLE		
	BERT-Large w/ MSP	**0.756**	**0.824**
	BERT-Large w/ Energy	**0.821**	0.772
	BERT-Large w/ MHLNB	0.934	0.311
	BERT-Large w/ Cosine	**0.861**	**0.508**
	VAN		
	BERT-Base w/ MSP	0.686	0.878
	BERT-Base w/ Energy	0.778	0.758
	BERT-Base w/ MHLNB	0.857	0.670
	BERT-Base w/ Cosine	0.749	0.874
	EXT		
	BERT-Base w/ MSP	0.738	0.832
	BERT-Base w/ Energy	0.855	**0.568**
	BERT-Base w/ MHLNB	0.957	0.213
	BERT-Base w/ Cosine	0.861	0.550
	BARLE		
	BERT-Base w/ MSP	**0.746**	**0.818**
	BERT-Base w/ Energy	**0.857**	0.580
	BERT-Base w/ MHLNB	**0.977**	**0.107**
	BERT-Base w/ Cosine	**0.945**	**0.244**

Table 8: OOD detection performance of Yahoo-AGNews-five with different model scales.

task performance. We examine the main task performance in Table 9 and Table 10. We use the classification accuracy (**ACC**) as the evaluation metric. Since VAN is the approach trained with only the cross entropy loss (without pseudo data and the contrastive objectives), it can be seen as the benchmark. The experimental results in both tables show that the main task performance of BARLE remains consistent with that of VAN. This analysis confirms that BARLE is capable of improving the background-shift OOD detection performance while maintaining the ID task performance.

D Full Ablation Study Results

We present the full ablation study results on all ID/OOD pairs in Table 11. The results consistently show that applying background-contrastive loss is more effective than semantics-contrastive loss in background-shift OOD detection, and applying both can further improve the performance.

Task	ID dataset	OOD dataset	Method	ID ACC ↑
Topic categorization	Yahoo!Answers	AGNews	VAN	0.829
			EXT	0.830
			BARLE	0.830
Sentiment classification	SST2	IMDB (T)	VAN	0.960
			EXT	0.955
			BARLE	0.954
		Amazon (T)	VAN	0.959
			EXT	0.954
			BARLE	0.946
	Amazon	IMDB (T)	VAN	0.956
			EXT	0.954
			BARLE	0.951
		SST2 (T)	VAN	0.955
			EXT	0.952
			BARLE	0.953
	Amazon-Books	Amazon-DVDs (T)	VAN	0.958
			EXT	0.960
			BARLE	0.950
		Amazon-Electronics (T)	VAN	0.953
			EXT	0.956
			BARLE	0.950
		Amazon-Kitchen (T)	VAN	0.954
			EXT	0.960
			BARLE	0.950

Table 9: Main task classification performance averaged across four scoring mechanisms based on RoBERTa-large.

Model	Method	ID ACC ↑
RoBERTa-Large	VAN	0.829
	EXT	0.830
	BARLE	0.830
RoBERTa-Base	VAN	0.801
	EXT	0.802
	BARLE	0.810
BERT-Large	VAN	0.810
	EXT	0.805
	BARLE	0.814
BERT-Base	VAN	0.782
	EXT	0.765
	BARLE	0.769

Table 10: Main task classification performance averaged across four scoring mechanisms on Yahoo-AGNews-five with different model scales.

Task	ID/OOD dataset pair	α/β	OOD metric		ID metric
			AUROC ↑	FAR95 ↓	ACC ↑
Topic categorization	Yahoo!Answers / AGNews	- / -	0.780	0.677	0.829
		+ / -	0.792	0.636	0.822
		- / +	0.843	0.553	0.842
		+ / +	**0.854**	**0.543**	0.830
Sentiment classification	Amazon-Books / Kitchen	- / -	0.695	0.922	0.954
		+ / -	0.719	**0.903**	0.950
		- / +	0.732	0.931	0.953
		+ / +	**0.739**	0.927	0.950
	Amazon-Books / Electronics	- / -	0.743	0.895	0.953
		+ / -	0.765	0.882	0.950
		- / +	0.779	**0.858**	0.953
		+ / +	**0.787**	0.902	0.950
	Amazon-Books / DVDs	- / -	0.549	0.955	0.958
		+ / -	0.574	0.929	0.945
		- / +	0.599	0.922	0.953
		+ / +	**0.612**	**0.899**	0.950
	SST2 / IMDB	- / -	0.877	0.459	0.960
		+ / -	0.899	0.438	0.956
		- / +	0.927	0.424	0.947
		+ / +	**0.949**	**0.380**	0.954
	SST2 / Amazon	- / -	0.865	0.487	0.959
		+ / -	0.880	0.574	0.956
		- / +	0.890	0.464	0.946
		+ / +	**0.908**	**0.428**	0.946
	Amazon / IMDB	- / -	0.568	0.948	0.956
		+ / -	0.641	0.918	0.949
		- / +	0.621	0.933	0.958
		+ / +	**0.687**	**0.831**	0.951
	Amazon / SST2	- / -	0.925	0.399	0.955
		+ / -	0.965	0.243	0.949
		- / +	0.977	0.106	0.958
		+ / +	**0.981**	**0.083**	0.953

Table 11: Full results of applying different contrastive loss components on all eight ID/OOD pairs. The "-" and "+" denote setting the corresponding parameters to zeros or not respectively. The results are averaged across four scoring mechanisms.