Evaluating the Practical Utility of Confidence-score based Techniques for Unsupervised Open-world Intent Classification

Sopan Khosla
AWS AI Labs, Amazon
sopankh@amazon.com

Rashmi Gangadharaiah
AWS AI Labs, Amazon
rgangad@amazon.com

Abstract

Open-world classification in dialog systems require models to detect open intents, while ensuring the quality of in-domain (ID) intent classification. In this work, we revisit methods that leverage distance-based statistics for *unsupervised* out-of-domain (OOD) detection. We show that despite their superior performance on threshold-independent metrics like AUROC on test-set, threshold values chosen based on the performance on a validation-set do not generalize well to the test-set, thus resulting in substantially lower performance on ID or OOD detection accuracy and F1-scores. Our analysis shows that this lack of generalizability can be successfully mitigated by setting aside a *hold-out* set from validation data for threshold selection (sometimes achieving relative gains as high as 100%). Extensive experiments on seven benchmark datasets show that this fix puts the performance of these methods at par with, or sometimes even better than, the current state-of-the-art OOD detection techniques.

1 Introduction

Open intent detection is of significant importance in practical dialog systems. Prior art (Zhang et al., 2021a) has shown that an intent classifier's performance degrades when it encounters examples of an unseen intent. Open-world classification (Fei and Liu, 2016) tries to mitigate this by not only correctly classifying data that appeared in training (ID), but also detecting examples that are not a part of any existing class (OOD). Schölkopf et al. (2001) and Tax and Duin (2004) use SVMs to find the decision boundary of each positive class (ID). Bendale and Boult (2016) leverage deep neural networks to learn representations that capture high-level semantic concepts. To detect OOD samples, Hendrycks and Gimpel (2017) use the softmax probability as the confidence score, where some negative samples are used for confidence threshold discovery. Other works (Zhou et al., 2021; Ren et al., 2021; Podolskiy et al., 2021; Zhan et al., 2021) use the distance between a new sample and the ID distributions to define their confidence scores. Whereas, Zhang et al. (2021a) learn an adaptive decision boundary (ADB) of each positive class by only using ID data and thus removing the dependence on a confidence-score completely.

Threshold-based OOD detection allows for more control, especially in scenarios where correctly predicting ID intents takes priority over detecting negatives or vice-versa. This has motivated researchers to evaluate confidence-based methods on threshold-independent metrics like Area Under ROC curve (AUROC) or Area Under PR curve (AUPR) on test-sets for an unbiased comparison. This is especially true for works on distance-based (e.g. Mahalanobis distance, Cosine similarity) confidence-scores (Zhan et al., 2021; Ren et al., 2021; Zhou et al., 2021), which seldom comment on the threshold selection criteria or the threshold-dependent performance of the underlying method and thus fail to reveal much about their practical utility.

In this work, we evaluate state-of-the-art approaches that use **d**istance-**b**ased **s**tatistics (DBS) to arrive at confidence-scores for Open-World Classification. Unlike previous works, we specifically focus on their performance on threshold-dependent metrics. We show that threshold values (δ) chosen based on the performance on the validation-set, used to tune the classifier, do not generalize well on the test-set. This results in poor test-set ID/OOD Accuracy and F1-scores as compared to confidence-score-independent techniques like ADB on multiple benchmark datasets. We analyse this lack of generalizability and propose the use of a hold-out set of ID samples from validation data for threshold selection. This fix improves the threshold-dependent performance of DBS approaches putting their test accuracy and F1-scores on ID/OOD detection at par with, or sometimes even better, than previously proposed open-classification techniques.

2 Methodology

We explore multiple state-of-the-art strategies for unsupervised open-world intent classification. The term *unsupervised* here refers to the absence of open-intent samples during training. We consider two approaches that leverage **l**ogit-**b**ased **s**tatistics (LBS) as their confidence-score (i.e. Maximum Softmax Probability and Energy), two DBS approaches (i.e. Mahalanobis distance and Cosine similarity), and Adaptive Decision Boundary (ADB) that does not rely on confidence-scores.

Maximum Softmax Probability (MSP). Several prior works adopt this method as a baseline for OOD detection (Hendrycks and Gimpel, 2017; Hsu et al., 2020; Hendrycks et al., 2020). MSP uses the maximum class probability $1 - max_{j=1}^{C}(p_j)$ among C training classes as its OOD indicator. p_j denotes the probability of j^{th} class.

Energy. Liu et al. (2020) show that energy scores not only better distinguish ID and OOD samples than softmax scores, but also align with the probability density of the inputs. A higher energy score indicates a higher likelihood of OODness.

Mahalanobis Distance (Maha) can be used to calculate the distance of an input sample to a distribution of samples from class c. We follow (Lee et al., 2018; Zhou et al., 2021) to compute the Mahalanobis distance from the penultimate layer of the transformer model by fitting a class-conditional multivariate Gaussian distribution. Finally, the OOD score for an instance is calculated as the minimum Mahalanobis distance among the C classes.

Cosine Similarity (Zhou et al., 2021)**.** The OOD score is calculated as the negative of the maximum cosine similarity between an instance at inference time and samples in the validation set.

Adaptive Decision Boundary (ADB) (Zhang et al., 2021a) does not rely on an OOD score for open-world classification. This approach aims to learn the euclidean distance decision boundaries for every seen class using the representations extracted from the pre-trained multi-class classification model trained on labeled ID training data. These spherical decision boundaries act as the distinction between ID and OOD samples.

Dataset	TRAIN-ID	VAL-ID	VAL-OOD	TEST-ID	TEST-OOD
CLINC	15,000	3,000	100	4,500	1,000
ROSTD	30,000	4,000	1,500	8,600	3,000
BANK77OOS	5,905	1,506	730	2,000	2,080
OOSBANK	500	500	600	500	1,350
OOSCREDIT	500	500	600	500	1,350
BANK	9,003	1,000	-	3,080	-
SO	12,000	2,000	-	6,000	-

Table 1: Data Statistics (SO = STACKOVERFLOW). -ID and -OOD refer to the in-domain and out-of-domain utterances present in each split.

3 Experimental Setup

3.1 Data

We evaluate the open-world intent classification strategies on six challenging benchmark datasets. Table 1 provides details on dataset statistics.

CLINC contains 150 intents, 22,500 ID queries and 1,200 OOD queries (Larson et al., 2019).

BANK includes 13,083 customer service queries across 77 intents in the banking domain (Casanueva et al., 2020).

STACKOVERFLOW (Xu et al., 2015) contains 20 different classes of technical question titles. BANK and STACKOVERFLOW do not contain explicit OOD utterances, so we follow (Shu et al., 2017; Zhang et al., 2021a) and only consider 75% samples from all the classes as seen classes.

ROSTD extends the English part of multilingual dialog dataset (Schuster et al., 2019) with OOD utterances. Following Gangal et al. (2020), we evaluate the different techniques on the variant with 12 fine-grained ID classes.

Zhang et al. (2021b) proposed two datasets. The first contains utterances from two domains, i.e., the "Banking" (**OOSBANK**) and "Credit cards" domain (**OOSCREDIT**) with both (1) out-of-domain and out-of-scope (OOD-OOS) queries and (2) in-domain but out-of-scope (ID-OOS) queries. The second dataset (**BANK77OOS**) extends BANK to include ID-OOS queries based on 27 held-out semantically similar in-scope intents. We combine both OOD-OOS and ID-OOS into a common OOD class.

3.2 Evaluation Metrics

We evaluate the performance of different open-world classification techniques on threshold-independent metrics like $AUROC$ and $AUPR_{out}$. Following previous work (Shu et al., 2017; Lin and Xu, 2019), we also evaluate the overall performance on accuracy (Acc) and macro F1-score on

	Performance on VAL (Pipeline 1) / VAL-HOLD (Pipeline 2)						Performance on TEST set (Pipeline 1 / Pipeline 2)					
	$AUROC \uparrow$	$AUPR_{out} \uparrow$	$F1_{All} \uparrow$	$F1_{In} \uparrow$	$F1_{Out} \uparrow$	$Acc \uparrow$	$AUROC \uparrow$	$AUPR_{out} \uparrow$	$F1_{All} \uparrow$	$F1_{In} \uparrow$	$F1_{Out} \uparrow$	$Acc \uparrow$
CLINC												
MSP	96.2 / 96.4	62.2 / 82.6	96.4 / 95.0	96.7 / 95.2	60.7 / 74.5	95.4 / 93.6	96.5 / 96.7	87.4 / 87.8	93.0 / 93.6	93.2 / 93.7	75.3 / 77.6	90.1 / 90.9
Energy	96.8 / 97.1	68.9 / 87.3	96.5 / 95.4	96.7 / 95.5	66.3 / 79.5	95.8 / 94.2	97.0 / 97.1	89.8 / 90.2	93.2 / 94.0	93.3 / 94.1	77.5 / 80.9	90.6 / 91.8
Cosine	100.0 / 98.1	100.0 / 88.7	97.2 / 95.4	97.2 / 95.5	100.0 / 80.9	97.0 / 94.5	97.4 / 97.4	90.1 / 90.1	53.8 / 94.1	53.9 / 94.2	43.9 / 81.5	52.3 / 91.8
Maha	99.7 / 98.3	98.2 / 89.6	97.4 / 95.6	97.6 / 95.7	80.8 / 83.3	97.0 / 94.8	97.6 / 97.6	90.9 / 90.8	87.9 / **94.2**	88.0 / 94.3	69.2 / 82.1	83.7 / **92.1**
ROSTD												
MSP	89.8 / 91.1	82.0 / 92.5	91.4 / 88.7	93.2 / 89.9	69.9 / 73.6	87.1 / 78.0	89.1 / 90.2	81.6 / 82.4	91.1 / 90.5	93.0 / 92.2	68.7 / 69.8	87.0 / 87.1
Energy	89.7 / 91.5	83.9 / 93.4	92.2 / 89.0	94.0 / 90.4	69.8 / 72.9	87.4 / 77.9	89.0 / 90.7	83.1 / 85.0	91.9 / 91.3	93.8 / 93.1	68.7 / 69.7	87.2 / 87.3
Cosine	100.0 / 99.5	100.0 / 99.6	97.8 / 96.7	97.6 / 96.7	100.0 / 96.7	99.0 / 96.5	99.5 / 99.4	98.5 / 98.4	59.2 / 95.6	58.7 / 95.7	64.4 / 94.2	69.3 / 96.8
Maha	99.9 / 99.6	99.8 / 99.6	97.8 / 97.1	97.7 / 97.1	99.5 / 97.1	99.0 / 96.9	99.6 / 99.5	98.8 / 98.7	86.7 / **95.7**	86.4 / 95.8	90.1 / 94.8	94.2 / **96.9**
BANK77OOS												
MSP	87.9 / 87.6	79.8 / 91.5	82.2 / 74.4	82.4 / 74.3	72.1 / 80.7	79.0 / 76.8	90.6 / 89.8	91.6 / 91.2	78.3 / 77.8	78.3 / 77.7	82.1 / 82.1	79.7 / 79.5
Energy	90.0 / 89.8	84.0 / 93.3	83.1 / 76.1	83.2 / 75.9	75.6 / 84.2	80.5 / 79.9	92.3 / 91.7	93.5 / 93.1	79.5 / 79.5	79.4 / 79.4	84.5 / 85.0	81.5 / 82.0
Cosine	100.0 / 91.8	100.0 / 94.2	89.9 / 77.5	89.7 / 77.3	100.0 / 86.7	93.0 / 82.3	93.5 / 93.6	94.1 / 94.1	7.3 / 80.0	6.0 / 79.9	68.3 / 86.7	52.5 / 83.1
Maha	99.3 / 92.3	99.3 / 94.7	89.5 / 77.7	89.4 / 77.5	96.5 / 87.4	91.6 / 82.9	94.2 / 94.1	94.9 / 94.7	57.8 / **80.1**	57.4 / 79.9	78.7 / 87.3	71.6 / **83.4**
OOSBANK												
MSP	90.0 / 90.0	92.3 / 95.6	85.9 / 81.9	86.5 / 81.6	80.4 / 84.8	81.0 / 80.8	93.5 / 93.8	97.2 / 97.3	83.3 / 83.5	82.6 / 82.7	90.6 / 91.9	86.8 / 88.2
Energy	88.6 / 88.8	92.0 / 95.4	85.7 / 79.5	86.4 / 79.2	78.9 / 82.7	80.1 / 78.5	93.3 / 93.9	97.5 / 97.7	83.4 / 82.0	82.7 / 81.1	90.3 / 91.2	86.4 / 87.5
Cosine	100.0 / 94.4	100.0 / 97.2	99.1 / 84.0	99.0 / 83.4	100.0 / 90.3	99.7 / 86.8	96.0 / 96.2	98.3 / 98.3	31.5 / **84.2**	25.9 / 83.2	86.8 / 93.7	77.7 / 90.7
Maha	100.0 / 94.6	100.0 / 97.4	99.1 / 84.7	99.0 / 84.1	100.0 / 91.0	99.7 / 87.8	96.6 / 96.6	98.6 / 98.6	20.7 / **84.2**	14.3 / 83.2	85.6 / 93.9	75.4 / **91.0**
OOSCREDIT												
softmax	89.1 / 90.8	90.6 / 95.4	83.1 / 80.3	83.4 / 79.7	80.4 / 86.3	80.9 / 82.3	93.4 / 94.1	97.0 / 97.2	81.2 / 82.7	80.3 / 81.8	90.0 / 91.9	86.4 / 88.7
energy	87.9 / 89.6	90.7 / 95.2	82.2 / 77.5	82.7 / 77.1	76.8 / 81.7	78.5 / 77.6	93.2 / 93.9	97.2 / 97.5	80.5 / 81.5	79.7 / 80.6	88.4 / 90.2	84.6 / 86.7
cosine	100.0 / 94.9	100.0 / 97.0	98.4 / 86.7	98.3 / 86.2	100.0 / 92.5	99.1 / 89.7	96.4 / 96.5	98.2 / 98.2	44.3 / 88.4	39.8 / 87.7	88.7 / 95.4	81.3 / 93.2
maha	100.0 / 95.4	100.0 / 97.4	98.4 / 87.6	98.3 / 87.0	100.0 / 93.3	99.1 / 90.7	97.2 / 97.1	98.7 / 98.7	61.1 / **88.8**	58.1 / 88.1	91.1 / 95.6	85.6 / **93.7**
BANK-75%												
MSP	88.2 / 89.2	71.3 / 74.8	88.6 / 88.0	89.0 / 88.3	66.0 / 66.0	83.1 / 83.1	86.7 / 87.1	69.7 / 69.9	87.8 / 87.5	88.2 / 87.9	64.5 / 63.1	82.2 / 81.6
Energy	88.2 / 89.4	73.5 / 78.0	88.9 / 88.1	89.3 / 88.5	66.5 / 69.6	83.4 / 84.0	86.5 / 86.8	71.5 / 71.5	87.9 / 87.2	88.3 / 87.6	65.8 / 66.7	82.5 / 82.4
Cosine	100.0 / 91.7	100.0 / 79.4	95.6 / 89.0	95.5 / 89.3	100.0 / 73.4	96.6 / 85.6	89.9 / 89.5	74.8 / 74.2	23.7 / **88.4**	23.3 / 88.7	43.6 / 69.9	36.3 / 83.6
Maha	100.0 / 92.2	100.0 / 80.1	95.6 / 89.4	95.5 / 89.6	100.0 / 77.3	96.6 / 86.5	90.6 / 90.4	74.8 / 74.9	37.8 / 87.9	37.7 / 88.2	47.1 / 72.1	44.3 / **83.7**
STACKOVERFLOW-75%												
MSP	90.0 / 90.1	68.5 / 68.3	86.7 / 85.9	87.7 / 86.8	71.8 / 71.3	83.1 / 82.8	90.0 / 90.5	68.5 / 69.3	86.7 / 86.8	87.7 / 87.8	71.5 / 71.8	83.1 / 83.4
Energy	90.7 / 90.8	69.6 / 69.2	87.3 / 86.5	88.2 / 87.4	73.4 / 72.8	84.0 / 83.5	90.6 / 91.2	69.6 / 70.4	87.1 / 87.2	88.1 / 88.2	72.9 / 73.3	83.7 / 84.0
Cosine	100.0 / 91.5	100.0 / 69.2	91.4 / 87.0	90.8 / 87.8	100.0 / 75.0	93.1 / 84.4	91.9 / 92.0	70.6 / 71.6	28.2 / **87.9**	27.1 / 88.7	45.9 / 75.7	39.9 / **84.9**
Maha	99.7 / 91.6	99.6 / 69.7	91.3 / 87.1	91.0 / 87.9	96.0 / 75.4	92.4 / 84.4	91.9 / 92.2	69.7 / 71.5	74.7 / **87.8**	75.4 / 88.6	63.9 / 75.7	71.8 / **84.9**

Table 2: OOD detection performance of confidence-score based techniques on different benchmark datasets (↑: higher is better). Test $F1_{All}$ and Acc scores for the best performing pipeline are underlined. Highest scores on the datasets are in **bold**.[1,2] Models that leverage distance-based scores (DBS; *Maha* and *Cosine*) and are trained using Pipeline 1 consistently perform poorly on threshold-depenedent metrics on the test-set. Furthermore DBS models that use Pipeline 2 substantially outperform their Pipeline 1 counterparts on all datasets (Columns 10-13; green).

known classes ($F1_{In}$), open class ($F1_{Out}$), and all classes combined ($F1_{All}$). The latter four metrics can only be calculated once a threshold is chosen.

3.3 Hyperparameters

We leverage the RoBERTa-base model implemented in the HuggingFace library for classification and use most of the default hyperparameters.[3] We experiment with training batch sizes $\{32, 64, 128\}$. Model with batch size 64 performs the best across all datasets. The learning rate for ID classifier training is set to 2e-5. [4]

3.4 Holdout set for threshold selection

Prior open-world classification research (Lin and Xu, 2019; Zhang et al., 2021a,b) uses the ID (VAL-ID) and OOD (VAL-OOD) samples in the validation data for threshold (δ) selection (**Pipeline 1**). We also experiment with a second setup that splits VAL-ID into two parts. VAL-TUNE-ID is used to tune the in-domain classifier, whereas the other (VAL-HOLD-ID), along with VAL-OOD[5], helps in deciding δ (**Pipeline 2**). For each dataset, we randomly sample one-third of VAL-ID as our VAL-HOLD-ID.

Following prior art (Zhang et al., 2020, 2021b), we tune δ to maximize $(A_{in} + R_{oos})$. A_{in} and R_{oos} represent the ID accuracy and the out-of-scope recall respectively on VAL / VAL-HOLD set.

4 Results and Analysis

Table 2 shows the performance of all compared methods on both pipelines. We report the averaged scores on 10 random seeds.[6]

[1] Each result is an average of 10 runs with different seeds.

[2] Scores on VAL cannot be compared to VAL-HOLD (columns 2-7).

[3] https://huggingface.co/roberta-base

[4] All experiments are run on a Tesla V100 16GB GPU.

[5] VAL-HOLD = VAL-HOLD-ID + VAL-OOD

[6] We exclude the std. dev. values due to lack of space.

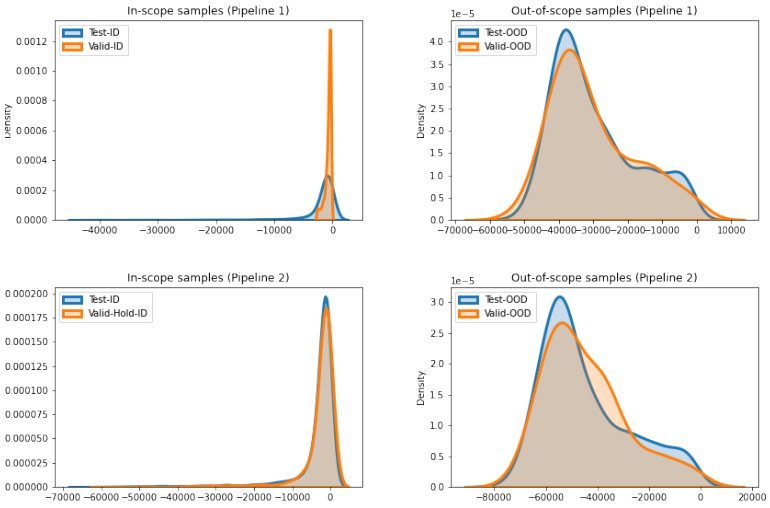

Figure 1: Maha distance (score) density plots for ID and OOD samples in CLINC dataset (VAL-OOD = VAL-HOLD-OOD). Top two charts show the density distribution for the model trained using Pipeline 1, whereas the bottom two focus on the model that uses Pipeline 2. We note that for Pipeline 1, the curve for VAL-ID looks substantially different from TEST-ID (top-left), suggesting that the thresholds selected using VAL-ID (Pipeline 1) might not generalize to the test set. Compare this to Pipeline 2 in-scope curve (bottom-left), where VAL-HOLD-ID almost exactly mimics the distribution of TEST-ID scores.

Dataset		$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc
CLINC	Maha	**94.2**	94.3	82.1	**92.1**
	ADB	93.3	93.4	79.3	90.6
	ADB-R	**94.3**	94.4	81.7	92.0
ROSTD	Maha	**95.7**	95.8	94.8	**96.9**
	ADB	95.0	95.7	86.5	93.3
	ADB-R	95.1	95.8	86.3	93.3
BANK77OOS	Maha	80.1	79.9	87.3	83.4
	ADB	78.6	78.5	84.7	81.6
	ADB-R	**81.1**	81.0	87.1	**83.9**
OOSBANK	Maha	**84.2**	83.2	93.9	**91.0**
	ADB	81.4	80.5	90.0	86.0
	ADB-R	81.9	81.1	89.5	85.5
OOSCREDIT	Maha	**88.8**	88.1	95.6	**93.7**
	ADB	82.8	82.0	90.8	87.2
	ADB-R	79.4	78.7	86.8	82.8
BANK-75%	Maha	87.9	88.2	72.1	**83.7**
	ADB*	86.0	86.3	66.5	81.1
	ADB-R	**88.4**	88.7	69.5	83.4
SO-75%	Maha	**87.8**	88.6	75.7	**84.9**
	ADB*	86.0	86.8	73.9	82.8
	ADB-R	87.6	88.5	74.5	84.3

Table 3: Test-set OOD detection performance of Cosine and Maha (Pipeline 2), and ADB variants on Accuracy and different F1-measures. ADB* denotes the official scores from (Zhang et al., 2021a). Maha (Pipeline 2) significantly outperforms ($p < 0.01$) ADB variants on ROSTD, OOSBANK, OOSCREDIT, and STACKOVERFLOW-75% datasets.

Models trained using Pipeline 1. In line with prior work (Zhou et al., 2021; Podolskiy et al., 2021), we find that Maha and Cosine perform better on the threshold-independent metrics ($AUROC$ and $AUPR_{out}$) across all datasets. This suggests that they are better at distinguishing ID instances from those considered to be OOD.[7]

Evaluation on threshold-dependent metrics (Acc and $F1$ scores) shows that the results obtained by MSP and Energy (LBS) on the test set do not differ much from the valid set, suggesting that the chosen δ generalizes well to unseen data. Compare this to Cosine and Maha (DBS) whose performance sees a drastic drop on the test set, despite achieving better scores on the valid set. This suggests that thresholds selected using Pipeline 1 for DBS might not transfer well to data in the wild, making them less useful in practice for OOD detection.

Models trained using Pipeline 2. On most datasets, the performance of these models on the test set mirrors that on the VAL-HOLD set. Furthermore, we see a consistent improvement in test Acc and $F1$ scores of all confidence-score methods as compared to their Pipeline 1 counterparts. Cosine and Maha see the highest gains, witnessing relative boosts as high as 100% on BANK-75% and STACKOVERFLOW-75%. Overall, thresholds chosen using Pipeline 2 seem to hold up better on unseen samples across the board, with Maha outperforming all other strategies on most datasets.

The top two plots in Figure 1 show the density plot of Mahalanobis distance values over CLINC ID and OOD data on VAL and test sets. We observe that although the distributions of TEST-OOD and VAL-OOD are quite similar, there are significant differences between the graphs for ID data (VAL-ID vs TEST-ID). There seem to be no VAL-ID samples with Maha score below -3000, whereas for TEST-ID, a substantial number of instances lie below -3000. This discrepancy might be a result of the slight overfitting of the trained ID classifier on VAL-ID samples as it leverages them for tuning. Compare this to the bottom two curves (in Figure 2) which plot Test vs VAL-HOLD instances. The density plots for both ID and OOD samples are almost identical.[8] Therefore, thresholds selected using VAL-HOLD are more likely to generalize to the unseen test set.

Comparison against ADB. ADB is the current state-of-the-art approach for unsupervised OOD detection. In Table 3, we report the performance of ADB (Zhang et al., 2021a)[9] and ADB-R where we replace the BERT encoder with RoBERTa-base

[7] Threshold-independent metrics cannot be calculated for ADB as it does not use a confidence-score for OOD detection.

[8] We see similar patterns across all datasets, but leave those figures out for brevity.

[9] https://github.com/thuiar/Adaptive-Decision-Boundary

and train the entire encoder during training. Maha (Pipeline 2) significantly outperforms ($p < 0.01$)[10] ADB and ADB-R on ROSTD, OOSBANK, OOS-CREDIT, and STACKOVERFLOW-75% while being competitive with the best performing ADB variant on the other three datasets.

5 Discussion and Conclusion

In this work, evaluate four confidence-score based unsupervised OOD detection techniques on seven state-of-the-art datasets. Most prior research (Zhou et al., 2021; Podolskiy et al., 2021) on methods that leverage distance-based statistics like Mahalanobis distance (Maha) or Cosine similarity (Cosine) only reports results on threshold-independent metrics like AUROC or AUPR. However, we show that despite their superior performance on AUROC, these techniques observe substantially lower scores on test ID and OOD detection Accuracy and F1-scores, when the entire validation-set (used to tune the ID classifier) is leveraged for threshold selection. This severely limits their practical utility.

Our analysis suggests that this discrepancy might be a result of the inadvertent overfitting of the trained classifier on VAL-ID samples. We show that this issue can be mitigated by leveraging a different evaluation setup that sets aside a hold-out set (not used during ID classifier tuning) from validation data for threshold selection. We observe that this new setup yields generalizable threshold values thus substantially improving the performance of Maha and Cosine on threshold-dependent metrics and making them more useful in real-world applications. Going forward, based on these findings, we would like to implore other researchers to also report the performance of their open-world classification approaches on threshold-dependent evaluation metrics, if applicable.

References

Abhijit Bendale and Terrance E Boult. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of ICLR*.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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)*, pages 1311–1316, Hong Kong, China. 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. *Advances in neural information processing systems*, 31.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496.

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.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

[10] We performed a one-tailed t-test to evaluate significance.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *Proceedings of ICML*.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805.

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916.

David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning*, 54(1):45–66.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. 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 3521–3532.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Jian-Guo Zhang, Kazuma Hashimoto, Yao Wan, Ye Liu, Caiming Xiong, and Philip S Yu. 2021b. Are pre-trained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. *arXiv preprint arXiv:2106.04564*.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.