PnPOOD : Out-Of-Distribution Detection for Text Classification via Plug and Play Data Augmentation

Mrinal Rawat [1] **Ramya Hebbalaguppe** [1] **Lovekesh Vig** [1]

Abstract

While Out-of-distribution (OOD) detection has been well explored in computer vision, there have been relatively few prior attempts in OOD detection for NLP classification. In this paper we argue that these prior attempts do not fully address the OOD problem and may suffer from data leakage and poor calibration of the resulting models. We present PnPOOD, a data augmentation technique to perform OOD detection via out-of-domain sample generation using the recently proposed Plug and Play Language Model(Dathathri et al., 2020). Our method generates high quality discriminative samples close to the class boundaries, resulting in accurate OOD detection at test time. We demonstrate that our model outperforms prior models on OOD sample detection, and exhibits lower calibration error on the 20 newsgroup text and Stanford Sentiment Treebank dataset (Lang, 1995; Socher et al., 2013). We further highlight an important data leakage issue with datasets used in prior attempts at OOD detection, and share results on a new dataset for OOD detection that does not suffer from the same problem.

1. Introduction

Most conversational agents deployed for enterprise applications have a specific purpose, such as assisting employees to answer questions about HR policies or resolving IT infrastructure issues. These agents are trained to answer queries with fixed intents from a particular domain via a text classifier that classifies user queries into one of several pre-defined intents. However, deep NLP models employed for intent classification in conversational systems are susceptible to improper responses to user queries due to overconfident predictions on out-of-domain (OD) test samples (Guo et al., 2017; Hendrycks & Gimpel, 2017). Automatic detection and redirection of OOD samples to other bots or for manual intervention would improve user experience and enhance trust in such systems. Prior attempts at OOD detection for text/image classification have varied from entropy maximization based methods (Lee et al., 2017; 2018), data augmentation techniques (Hein et al., 2019; Hendrycks & Dietterich, 2019; Patel et al., 2021) and uncertainty quantification methods(Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). However, often these methods ignore the impact of OOD detection on the calibration error of the intent classifier. A further problem with OOD detection datasets reported in prior work is that the OOD samples generated for detection training and those encountered during testing may belong to overlapping domains. To this end, we present PnPOOD, an OOD detection algorithm that significantly outperforms prior methods, both on OOD detection metrics and model calibration error on text classification. The technique utilizes a PPLM model (Dathathri et al., 2020), a computationally light technique to guide sentence generation for OOD samples, which are then used to train the OOD classifier. To ensure high quality OOD sample generation, we provide initial guiding tokens from in-domain (ID) samples. We demonstrate the superiority of our technique on the SST dataset (Socher et al., 2013), on which all prior results have been reported in the recent past(Hendrycks & Dietterich, 2019; Hendrycks & Gimpel, 2017). Further, we discovered a data leakage issue with the previous approaches that has so far been overlooked. Therefore, in order to provide an unbiased assessment of our model, we also conducted experiments on a dataset carved out of the 20 newsgroup(Lang, 1995) dataset ensuring no data leakage between train and test OOD samples.

The primary contributions of this work are: (a) We propose PnPOOD, a new technique for generating and training OOD detectors on OOD samples that are close to the class decision boundaries. (b) Results show our technique outperforms all existing methods for OOD detection on text classification in the recent past, both in terms of detection accuracy and model calibration

[*]Equal contribution [1]TCS Research & Innovation. Correspondence to: Mrinal Rawat <rawat.mrinal@tcs.com>, Ramya Hebbalaguppe <ramya.hebbalaguppe@tcs.com>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

2. Related Work

The recent techniques fine-tune hyper-parameters on a validation set to optimize OOD detection, for example, Hendrycks and Gimpel (MSP) (Hendrycks & Gimpel, 2017) use the maximum confidence scores from a Softmax output as a detection score, which in turn is used to classify OOD samples. ODIN utilizes temperature scaling with input perturbations using the OOD validation dataset to tune hyper-parameters (Liang et al., 2017). However, hyper-parameters tuned with one OOD dataset is found not to generalize to other datasets. The sorted Euclidean distance between the input and the k-nearest training samples has also been used as a detection score (Zhang et al., 2006). The likelihood ratio method for deep generative model corrects for confounding background statistics and is used as an effective OOD detection method in image classifiers. A background model is trained using perturbed IND samples employing a Liklihood ratio which enhances OOD detection performance(Ren et al., 2019). Lee (Lee et al., 2018) propose detecting OOD samples by training a logistic regression detector on the Mahalanobis distance vectors calculated between test images' feature representations and the class conditional Gaussian distribution at each layer. (Lee et al., 2017) generate synthetic images sampled from the low density boundary regions of the in-distribution space. For generating synthetic OOD samples around uniform distribution, they propose using GANs. OOD samples can be forced to have an uniform distribution by minimizing the Kullback-Leibler divergence between the model generated probabilities on OOD samples and a uniform distribution. The OOD detectors listed before are demonstrated mainly on visual tasks, in particular, on image classification. Our focus has been to advance the state-of-the-art in OOD detection in text classification where the research efforts are scant and our technique is focussed to intent classification in natural language processing using Natural language generation that we describe in future.

3. Methodology

For formulation/definition of OOD detection in a supervised classification setting, See (Liang et al., 2017). We propose a practical approach to post-hoc OOD detection i.e. situations where OOD detection has to be incorporated into an existing classification model without model retraining. The dataset D comprises of sentences from N different domains. Out of these N domains, k domains are treated as IN-DISTRIBUTION (IND) while the rest of the $N-k$ domains are treated as OUT-DISTRIBUTION (OOD). Given a set of IND sentences $D_{ind} = (x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$, we train a network using the following loss function:

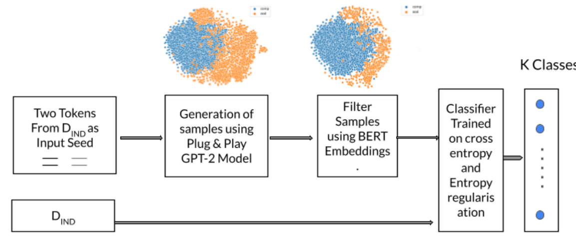


Figure 1. An OOD classifier is trained on OOD samples generated using Plug and Play Language model and IND data. At a high level we employ entropy regularisation on OOD samples. Intuitively, when we force the samples to have highest entropy, OOD samples are closer to the uniform distribution (zero confidence). A sample is considered OOD based on observing the probability vector at the classifiers output being low on all the elements of a prediction vector. Also, our technique filters the subset of OOD samples generated from PPLM to concentrate around the class boundary thereby improving the OOD performance.

$$\mathcal{L} = \underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(x,y)\sim D_{\text{ind}}^{\text{train}}}[\mathcal{L}_{\text{CE}}(y_{in}, f_\theta(x))] \\ +\alpha \cdot \mathbb{E}_{(x_{ood}^{PPLM})\sim D_{\text{out}}^{\text{OOD}}}[\mathcal{L}_{\mathcal{E}}(f_\theta(x_{ood}^{PPLM}), U(y))] \qquad (1)$$

where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss, $f_\theta(x)$ denotes the softmax output prediction of neural network for an input sample x. We obtain least loss when $\alpha = 1$ is used our experiments. x_{ood}^{PPLM} is the OOD sample generated from our data augmentation using PPLM light weight models as detailed in Fig 1. $D_{\text{out}}^{\text{OOD}}$ is the OOD dataset. $\mathcal{L}_{\mathcal{E}}$ is a regularisation that tries to minimize the loss between output probability vector of an OOD sample x_{ood}^{PPLM} to an uniform distribution inspired from (Lee et al., 2017). Note that it is harder to model a complete test distribution as the OOD space is infinite. However, we can use heuristics for detecting test distribution using just the representations from only IND data. Our approach is one such that considers taking the low density samples at the class boundaries of IND distribution to model OOD. Section 3.1 throws light on OOD sample generation to obtain $D_{\text{out}}^{\text{OOD}}$ for training an OOD detector as explained in Equation 1. We notice, that the training procedure using our OOD samples significantly helps in boosting the OOD detection performance.

3.1. PPLM based OOD data augmentation

We propose a novel technique of generating the sentences which are close to the IN-Domain samples. We hypothesize that the samples which are closer in the embedding space to the IND cluster boundary are more relevant for discriminating between IND and OOD samples. This is particularly important as the OOD space is huge and there exists no principled way of generating samples that are representative of the entire OOD space. These generated sentences form proxy for OOD which is used for entropy regularisation, thus improving the OOD detection performance. Fig 1 depicts the OOD sample generation and training procedure.

While large scale language models like GPT-2 (Radford et al., 2019) have shown remarkable performances in Natural Language generation (NLG), guiding the generation for a task often required compute intensive fine tuning of the full model. The recent proposed Plug and Play Language Models (PPLM) (Dathathri et al., 2020) utilize an attribute model in order to guide generation without fine tuning the large model. In this work, we utilize PPLM to generate new samples whose BERT(Devlin et al., 2018) embeddings are closer to IN-DOMAIN cluster boundary. To generate the samples using the PPLM, we provide the initial input seed and the bag-of-words which is used to control the generation. We take the initial two tokens from random samples in D_{ind} as input seed and generate sentences directed towards D_{ood} using the light weight PPLM training procedure. The bag-of-words are extracted from the out-domain dataset since they guide the model to generate the sentences which are out-of-distribution. For example,

Original Sentence: "This article includes answers what options have for software intel based unix system"

Input: [Science] <This article>

Output: This article explores a recent study on a large scale of global climate system climate change, which finds no direct evidence of the Earth's climate warming.

Words from D_{ood} (from science domain):

`astronomy, atom, biology, cell, chemical, chemistry, earth, climate`

Despite controlled generation, we observed that some generated samples were quite far from the IND samples. To overcome this problem, we filter the sentences based on the BERT embeddings (Wolf et al., 2020). We first find the IND cluster center C_{ind}. We then measure the distance of each sentence embedding with the cluster center and consider only the ones which are closer to the cluster boundary. Refer Alg. 1 in Supplemental material for pseudocode of our proposed OOD sample generation including OOD sample filtering.

3.2. Post-hoc calibration

The modern DNNs produce overconfident decisions, to overcome this, in addition to OOD detection, we apply Dirichlet calibration(Kull et al., 2019) as a post-hoc calibration technique. They assume $\hat{p}(X|Y=j) \sim \text{Dirichlet}(\alpha^{(j)})$, where $\alpha^{(j)} \in \mathbb{R}^K$. They propose a new regularisation method called Off-Diagonal regularisation, given by $ODIR = \frac{1}{k*(k-1)} \sum_{i \neq j} w_{i,j}^2$. This post-hoc calibration improves the consonance of predicted probabilities with the accuracies produced by the classifier. Dirichlet calibration can be thought of log-transforming the uncalibrated probabilities, followed by one linear layer and softmax; this simple procedure is known to outperform temperature/vector scaling to produce well-calibrated scores (Hinton et al., 2015).

4. Experiments

SST We evaluated our approach by training with SST (Socher et al., 2013) dataset, and tested on Multi30K(Elliott et al., 2016) and SNLI(Bowman et al., 2015) OOD datasets. The Stanford Sentiment Treebank (SST) dataset consists of movie reviews expressing positive or negative sentiment. SNLI is a dataset of predicates and hypotheses for natural language inference and Multi30K is a dataset of English-German image descriptions.

20 Newsgroups We evaluate our approach and compared against other baselines on The 20 Newsgroups (Lang, 1995) dataset. This dataset consists of approximately 20000 newsgroups documents divided into 20 groups. These 20 newsgroups correspond to different topics but some of the overlapping topics can be merged yielding 6 major domains. Table 2 illustrates domain and group information present in Appendix A.

Out of the four domains, we choose three `Computer, Sports and Politics` to train the system leaving out the 'misc' domain due to data leakage issues. We train the system on one domain considering it as ID dataset and test on the remaining domains as OOD datasets. For e.g. We train the system with D_{ind} = Computer and D_{ood} as Sports, Politics again to overcome issues of data leakage among domains. In this way, we experiment with all the possible combinations. We evaluate the system on the baseline approaches detailed in Supplemental material: (a) **Maximum Softmax Probability (MSP)(Hendrycks & Gimpel, 2017):** The maximum softmax score is used as a detection score based on the threshold. (b) **MSP + Entropy Reg. (ER)(Hendrycks et al., 2019):** A model is trained with the entropy term along with the cross entropy loss as described in Eq. 1. The detection score is calculated using the MSP. (c) **MSP + ER + PPLM (Our approach-PnPOOD):** A model is trained in a similar manner as in MSP + ER. The samples provided as OOD are generated using the PPLM as described in Section 3.1. The detection score is calculated using the MSP.

For outlier exposure, we need the train the system with some OOD samples. However, a glaring limitation of the previous state-of-the-art approaches is that they use a generic dataset like Gutenberg(Lahiri, 2014) derived from a large collection of books for obtaining OOD samples, and some domains from this dataset may overlap with ID samples resulting

Table 1. Evaluation of OOD detection performance on modified **20newsgroup dataset** as described in Suppl section A. Note the last column is ECE + Direchlet calibration(Kull et al., 2019). Note, our proposed method, PnPOOD demonstrates best OOD detection performance and the least model calibration error. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages. **Bold** numbers are superior results.

DATASET (IND)	DATASET (OOD)	METHOD	FPR@90 ↓	AUROC ↑	AUPR ↑	ECE ↓	ECE (+ DIR. CAL.) ↓
COMPUTER	SPORTS	MSP(HENDRYCKS & GIMPEL, 2017)	0.72	0.62	0.23	0.56	0.41
		MSP + ER(HENDRYCKS ET AL., 2019)	0.26	0.9	0.64	0.33	0.28
		MSP + ER + PPLM (PNPOOD)	**0.18**	**0.92**	**0.65**	**0.32**	**0.26**
	POLITICS	MSP(HENDRYCKS & GIMPEL, 2017)	0.72	0.63	0.24	0.56	0.42
		MSP + ER(HENDRYCKS ET AL., 2019)	0.15	0.92	0.67	0.32	0.287
		MSP + ER + PPLM (PNPOOD)	**0.11**	**0.93**	**0.68**	**0.31**	**0.27**
SPORTS	COMPUTER	MSP(HENDRYCKS & GIMPEL, 2017)	0.71	0.63	0.23	0.73	0.6
		MSP + ER(HENDRYCKS ET AL., 2019)	0.32	0.82	0.35	0.4	0.33
		MSP + ER + PPLM (PNPOOD)	**0.22**	**0.89**	**0.51**	**0.39**	**0.31**
	POLITICS	MSP(HENDRYCKS & GIMPEL, 2017)	0.76	0.61	0.21	0.73	0.6
		MSP + ER(HENDRYCKS ET AL., 2019)	0.3	0.82	0.36	0.38	0.33
		MSP + ER + PPLM (PNPOOD)	**0.24**	**0.87**	**0.51**	**0.38**	**0.32**
POLITICS	COMPUTER	MSP(HENDRYCKS & GIMPEL, 2017)	0.61	0.67	0.25	0.72	0.61
		MSP + ER(HENDRYCKS ET AL., 2019)	0.24	0.91	0.64	0.48	0.41
		MSP + ER + PPLM (PNPOOD)	**0.2**	**0.92**	**0.6**	**0.45**	**0.39**
	SPORTS	MSP(HENDRYCKS & GIMPEL, 2017)	0.63	0.67	0.25	0.71	0.62
		MSP + ER(HENDRYCKS ET AL., 2019)	0.42	0.85	0.53	0.47	0.41
		MSP + ER + PPLM (PNPOOD)	**0.34**	**0.88**	**0.56**	**0.46**	**0.4**

in data leakage and misleading performance at test time. To address this the effectiveness of the approach, we use two other domains from the 20Newgroup dataset Science, Religion and consider their samples as D_{ood}^{OE} ((Hendrycks et al., 2019)). In our approach, instead of using D_{ood}^{OE} drawn from the 20Newsgroup dataset, we generate the samples D_{ER}^{PPLM} using the approach described in Section 3.

4.1. Evaluation Metrics

We evaluate our approach for OOD detection employing the standard metrics proposed in OOD literature such as AUROC, AUPR, ECE and FPR@90. We treat the OOD samples as the positive class as proposed by (Hendrycks & Gimpel, 2017). Refer Appendix A on details of evaluation metrics.

5. Results

We evaluated our approach on the metrics described in Section 4. We present the results on the OOD detection performance when D_{ind} is SST dataset and D_{ood} is MULTI30K, SNLI respectively. Our results demonstrate that our approach outperforms the other state-of-the-art approaches by a significant margin across all the metrics. We want to highlight that previous approaches have exploited the generic dataset like Gutenberg while training with entropy regularization. We believe that D_{ood} should be drawn from a different domain to avoid data leakage. Detailed description and results are present in Appendix C (See Tables on AUROC, AUPR, FPR@90TPR, ECE, ECE+Direchlet Calibration on the benchmark datasets).

To address the data leakage issues, we experiment with 20Newsgroup dataset. Table 1 illustrates the comparison of our approach with the previous state-of-the-art approaches. Our method outperforms all the other baselines with a significant margins in all the domains. We also report the Expected Calibration Error (ECE) before and after posthoc Dirichlet calibration.

6. Conclusion

In the paper, we proposed PnPOOD, a novel technique to detect the out-of-distribution samples using the pseudo OOD samples generated with the Plug and Play language model. We generate high quality samples which are closer to the IND sample cluster boundary, thus helping in the improving classification performance for OOD detection. In the future, we want to evaluate our approach on CLINC dataset (Larson et al., 2019; Liu et al., 2020) and other NLP tasks like token classification in sequence-to-sequence tasks.

References

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. Plug and play language models: A simple approach to controlled text generation. *ICLR 2020*, abs/1912.02164, 2020. URL https://openreview.net/pdf?id=H1edEyBKDS.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3210. URL http://www.aclweb.org/anthology/W16-3210.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

Hein, M., Andriushchenko, M., and Bitterwolf, J. 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*, pp. 41–50, 2019.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations (ICLR)*, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Kull, M., Perello-Nieto, M., Kängsepp, M., Song, H., Flach, P., et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

Lahiri, S. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 96–105, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E14-3011.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Lang, K. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., and Mars, J. An evaluation dataset for intent classification and out-of-scope prediction. *CoRR*, abs/1909.02027, 2019. URL http://arxiv.org/abs/1909.02027.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 7167–7177, Red Hook, NY, USA, 2018. Curran Associates Inc.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *CoRR*, abs/2006.10108, 2020. URL https://arxiv.org/abs/2006.10108.

Patel, K., Beluch, W., Zhang, D., Pfeiffer, M., and Yang, B. On-manifold adversarial data augmentation improves uncertainty calibration. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8029–8036. IEEE, 2021.

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

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October 2020. Association for Computational Linguistics.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2126–2136. IEEE, 2006.

A. Evaluation metrics

The following are description on the OOD detection metrics:

- **AUROC:** The area under the receiver operating characteristic (AUROC) is a common metric used in OOD detection, specifically, it is the area under the FPR against TPR curve. Higher value represents a better model.
- **AUPR:** The area under the precision-recall curve is used. Higher value represents better model. Particularly useful incase of ID-OOD sample imbalance.
- **FPR@90:** False positive rate (FPR) at 90% true positive rate. Lower is better.
- **ECE:** The Expected Calibration Error (ECE) takes a weighted average over the difference of absolute accuracy and confidence/prediction probability. Lower is preferred (Guo et al., 2017).

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|, \tag{2}$$

where n_b is the number of predictions in bin b, N is the total number of data points, and acc(b) and conf(b) are the accuracy and confidence of bin b, respectively.

B. Dataset description of 20newsgroup

Table 2 illustrates the detailed information about the domains and classes present in 20Newsgroup dataset. We want to highlight that we leave out Domain : 4 (misc) as it includes samples from other domains resulting in data leakage.

Table 2. Dataset Description of 20newsgroup

DOMAIN 1: COMPUTER	DOMAIN 2: SPORTS	DOMAIN 3: SCIENCE
COMP.GRAPHICS COMP.OS.MS-WINDOWS.MISC COMP.SYS.IBM.PC.HARDWARE COMP.SYS.MAC.HARDWARE COMP.WINDOWS.X	REC.AUTOS REC.MOTORCYCLES REC.SPORT.BASEBALL REC.SPORT.HOCKEY	SCI.CRYPT SCI.ELECTRONICS SCI.MED SCI.SPACE
DOMAIN 4: MISC.	**DOMAIN 5: POLITICS**	**DOMAIN 6: RELIGION**
MISC.FORSALE	TALK.POLITICS.MISC TALK.POLITICS.GUNS TALK.POLITICS.MIDEAST	TALK.RELIGION.MISC ALT.ATHEISM SOC.RELIGION.CHRISTIAN

C. Setup and Results

C.1. Configuration Details

We conduct experiment with a LSTM based text classifier. We stacked two LSTM layers of dimension 128. We initialize the input embeddings with pre-trained GloVe[1] vectors of size 300. We used the Adam optimizer with learning rate as 0.001 and drop out set to 0.3. The batch size is 32. We perform our experiments on NVidia GTX 1070 with 16 GB RAM. All the models are implemented in Pytorch.

C.2. Results

Table 3 demonstrates the results on the SST dataset. Our method outperforms previous state-of-the-art approaches by a significant margin.

Figure 3 shows the AUROC plot of sports domain as D_{ind} when D_{ood} is computer, politics respectively. Observe our proposed approach yields a higher AUROC in comparison to competing methods after Direchlet Calibration.

Figure 2 illustrates the t-SNE plot of our method w.r.t. to other baselines. We demonstrate that the samples generated using our approach yields a better OOD classification performance visually as depicted in t-SNE plot in Suppl. material.

Table 4 illustrates the in-distribution accuracy on 20Newsgroups dataset. We report results in comparison with previous baselines. We observe that including the entropy regularization did not affect the in-distribution classifier's performance.

D. Example success and failure cases

We train the classifier with D_{ind} as Computer and tested on the samples from D_{ood} as Sports. The following examples shows the success and failure cases:

Success Case:

Test data: "Headlights problem thanks all you who responded posting the problem with truck headlights low beam problem was loose wire connection was not the fuse minority you suggested thanks again"

Our model was able to give lower score to the OOD test sample input to the classifier. Refer Table 5.

Failure Case:

Text: "how hard change springs truck article apr michael apple com ems michael apple com michael smith writes does take any peculiar tools remove the rear springs from ford truck naah just coupla nice big bumps"

On closer inspection, we believe that the words like "apple", ".com" are from computer domain and may have confused the classifier. Refer Table 6

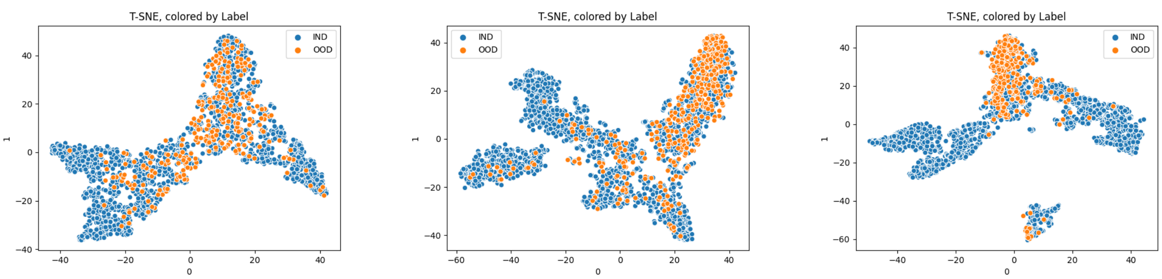
E. Algorithm for OOD sample generation

We generate OOD samples using PPLM (Dathathri et al., 2020) that utilize an attribute model in order to guide gener-

[1]https://nlp.stanford.edu/projects/glove/

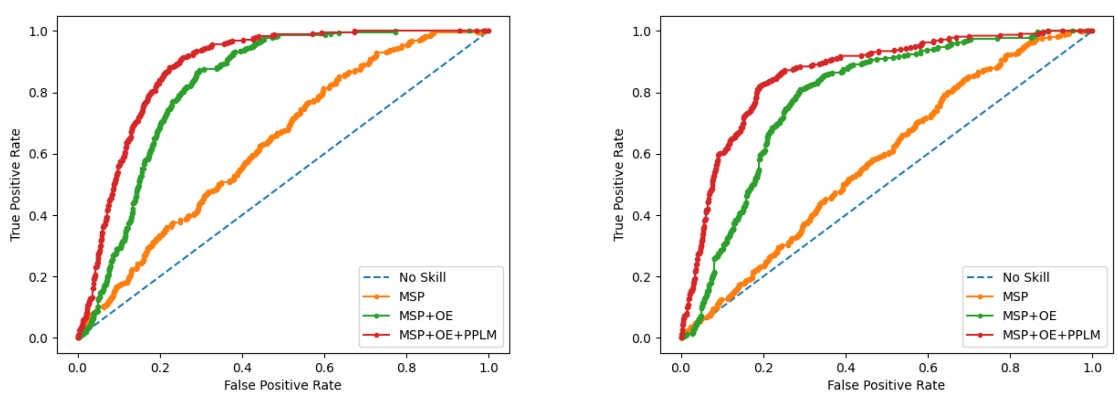
Table 3. Evaluation of OOD detection performance on **SST dataset** as in-distribution. Note, our proposed method, PnPOOD demonstrates best OOD detection performance and the least model calibration error. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages. **Bold** numbers are superior results.

DATASET (IND)	DATASET (OOD)	METHOD	FPR@90 ↓	AUROC ↑	AUPR ↑
SST	MULTI30K	MSP(HENDRYCKS & GIMPEL, 2017)	0.85	0.54	0.19
		MSP + ER(HENDRYCKS ET AL., 2019)	0.47	0.81	0.44
		MSP + ER + PPLM (PNPOOD)	**0.40**	**0.84**	**0.48**
	SNLI	MSP(HENDRYCKS & GIMPEL, 2017)	0.63	0.73	0.32
		MSP + ER(HENDRYCKS ET AL., 2019)	0.11	0.95	0.71
		MSP + ER + PPLM (PNPOOD)	**0.05**	**0.97**	**0.77**



(a) MSP (b) MSP + ER (c) MSP + ER + PPLM (PnPOOD)

Figure 2. t-SNE plot illustrating that our proposed approach yields better separation between IND and OOD samples on a pair of domains in 20newsgroup dataset.



(a) D_{ind} = Sports & D_{ood} = Computer (b) D_{ind} = Sports & D_{ood} = Politics

Figure 3. AUROC on different pairs of domains of 20Newsgroup, (a) AUROC when $D_{ind} = Sports$ and D_{ood} = Computer (b) AUROC when $D_{ind} = Sports$ and D_{ood} = Politics. Note: AUROC is a plot of False Positive Rate (x-axis) and True Positive Rate (y-axis), it shows the performance of classification models under all thresholds.

ation without fine tuning the large Natural language generation model. We also perform filtering of samples which are closer in the embedding space to the IND cluster boundary are more relevant for discriminating between IND and OOD samples. We conducted multiple experiments setting the threshold in the [0,5,10,12,20,25] and setting a threshold value as 10, we obtain the maximum OOD performance. We observed that increasing the threshold value greater than

Table 4. Evaluation of IND classifier accuracy on **20Newsgroups dataset**. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages. **Bold** numbers are superior results.

DATASET	METHOD	ACC ↑
COMPUTER	MSP(HENDRYCKS & GIMPEL, 2017)	0.46
	MSP + ER(HENDRYCKS ET AL., 2019)	0.47
	MSP + ER + PPLM (PNPOOD)	**0.48**
SPORTS	MSP(HENDRYCKS & GIMPEL, 2017)	0.73
	MSP + ER(HENDRYCKS ET AL., 2019)	0.73
	MSP + ER + PPLM (PNPOOD)	**0.74**
POLITICS	MSP(HENDRYCKS & GIMPEL, 2017)	**0.65**
	MSP + ER(HENDRYCKS ET AL., 2019)	0.64
	MSP + ER + PPLM (PNPOOD)	0.64

Table 5. Softmax scores

METHOD	SOFTMAX SCORE ↓
MSP(HENDRYCKS & GIMPEL, 2017)R	0.59
MSP + ER(HENDRYCKS ET AL., 2019)	0.26
MSP + ER + PPLM (PNPOOD)	**0.19**

Table 6. Softmax scores

METHOD	SOFTMAX SCORE ↓
MSP(HENDRYCKS & GIMPEL, 2017)	**0.65**
MSP + ER(HENDRYCKS ET AL., 2019)	0.84
MSP + ER + PPLM (PNPOOD)	0.98

10 increased the number of generated sentences and we observed the OOD performance reduced with increase in threshold. Similarly, by decreasing the threshold we were losing out on important OOD sentences.

Algorithm 1 Filtering the D_{ood}^{PPLM} samples

Inputs:
$\mathcal{D}_{ood}^{PPLM}$ `// OOD Samples`
$\mathcal{D}_{ind}$ `// IND Samples`
$\Gamma = 10$`// Threshold Distance`

Function `ClusterDistance(`E`)`
  $C \leftarrow \frac{1}{N}\sum_{i=1}^{N} E_i$; `// finds cluster center of embeddings`
  $Dist \leftarrow \emptyset$
  foreach e *in* E **do**
    $Dist_j \leftarrow EucDistance(C, e)$; `// Euclidean distance`
  $d = percentile(Dist, 0.95)$ `// remove outliers`
  return C, d

`// Training`
$E_{ood} \leftarrow BertEmbeddings(D_{ood}^{PPLM})$
$E_{ind} \leftarrow BertEmbeddings(D_{ind})$
$C, d \leftarrow$ `ClusterDistance(`E_{ind}`)` ; `// Distance from cluster center to the boundary`
$S \leftarrow \emptyset$; `// Filtered Sentences`
foreach s_i *in* $\mathcal{D}_{ood}^{PPLM}$ **do**
  if $d < EucDistance(C, s_i) < d + \Gamma$ **then**
    $S_j \leftarrow s_i$;
return S