OOD detection on text classification

***Andres GARCIA**
MS DS Student
https://github.com/andresgarciaparrado
andresfernando.garciaparrado@ensae.fr

***Aurelie NGALULA**
MS DS Student
https://github.com/AurelieNgalula
aurelie.ngalulangassam@ensae.fr

Abstract

In natural language processing (NLP) tasks, it is crucial to detect whether a given input is out-of-distribution (OOD), meaning it falls outside the model's training data. This is necessary because when the model encounters new text that it has not been trained on, it may not perform well and make incorrect predictions. The issue is particularly significant in scenarios where the model's predictions have real-world consequences, such as in medical diagnosis or financial fraud detection.

To address this problem, researchers have developed a baseline of three basic approaches for detecting OOD samples, which we have presented in our project available on GitHub[1]. We have evaluated the performance of these approaches in binary sentiment classification. Although they are effective in identifying OOD samples, there is still room for improvement in OOD detection methods, especially in correctly identifying true positive OOD samples. These techniques can serve as a starting point for practicioners who aim to apply OOD detection in NLP applications.

1 Introduction

Natural Language Processing (NLP) has made remarkable progress in recent years and has been implemented in various domains such as chatbots and sentiment analysis. However, these advancements have also raised concerns regarding the safety and fairness of NLP models. The issues that are particularly important in this regard include fairness, generalization to out-of-distribution (OOD) data (Gomes et al.), and adversarial defense.

Fairness is a crucial element in ensuring the safety of NLP models as models trained on biased data can result in discriminatory outcomes (Colombo, 2021; Pichler et al., 2022; Colombo et al., 2022b). To address this issue, researchers have suggested several strategies, such as carefully selecting datasets (Fabris et al., 2022), designing appropriate losses (Colombo et al., 2021b,a), and using post-processing techniques (Petersen et al., 2021). These approaches aim to eliminate biases related to attributes like age, gender, and race.

Another critical aspect of NLP safety is the ability of models to generalize to OOD data (Darrin et al., 2023a,b). Such data differs significantly from the training data, and the inability of the models to handle OOD data can lead to erroneous predictions. In domains like healthcare or legal decision-making, such inaccuracies can have dire consequences. Therefore, there is a growing interest in developing techniques that enhance the models' ability to generalize to OOD data.

Lastly, adversarial defense is a vital component of safety (Picot et al., 2023a,b). Adversarial attacks are deliberate attempts to manipulate input data to produce incorrect results from the model. To safeguard NLP models against such attacks, researchers have proposed several adversarial defense techniques to ensure that the models remain secure and robust.

Problem Framing

Let us define the following notations: X as the input space consisting of textual inputs, $Y = \{0, 1\}$ as the label space indicating binary output, and $D_n = \{(x_1, y_1), ..., (x_n, y_n)\} iid \sim p_{XY}$ as the sample data, where p_{XY} represents the probability distribution defined over $X \times Y$. The classifier trained on D_n is denoted by $f_n : X \to Y$.

The OOD classification problem involves a random variable $z \in \{0, 1\}$, where $z = 0$ if (x, y) belongs to the training distribution (IN), and $z = 1$ otherwise (ODD). The objective is to construct a similarity function $s : X \to R^+$ that can classify

[1] https://github.com/andresgarciaparrado/OOD_NLP

any test input x as IN or OOD using a given fixed threshold γ.

Specifically, if $s(x) > \gamma$, the input x is classified as IN ($\hat{z} = 0$), and if $s(x) \leq \gamma$, it is classified as OOD ($\hat{z} = 1$). This approach helps to differentiate between in-distribution and out-of-distribution samples (Colombo et al., 2022a).

The performance of OOD methods is evaluated using four metrics:

Area Under the ROC curve (AUROC) $\gamma \rightarrow (Pr(s(x) > \gamma | z = 0), Pr(s(x) \leq \gamma | z = 1))$ (Bradley, 1997)

Area Under the PR curve (AUPR) $\gamma \rightarrow (Pr(z = 1 | s(x) \leq \gamma), Pr(s(x) \leq \gamma | z = 1))$ (Davis and Goadrich, 2006)

False Positive Rate at 95% **True Positive Rate** (FPR)

Error of the best classifier (Err (%)), is calculated by selecting the threshold that yields the lowest classification error. In this case, we assume that the threshold with the best F1-score is the optimal choice.

2 Experiments Protocol

2.1 Data sets

We download the SST2 (Socher et al., 2013) and IMDB (Maas et al., 2011) datasets from Hugging Face's datasets library in Python. The selection of SST2 as the in-distribution dataset (IN-DS) and IMDB as the out-of-distribution dataset (OOD-DS) can be classified as an example of background shift for OOD, according to (Arora et al., 2021).

The text data is preprocessed by removing any HTML tags, non-alphanumeric characters, and stop words present in the data before further encoding. We split the IN-DS into training and test sets using a 80/20 split (Gholamy et al., 2018) and use 10% of the training set as validation. We randomly select a subset of examples from the IMDB dataset to use as OOD examples, ensuring that the proportion of the test set that should be OOD is 30%. Please refer to Figure 1 below.

Dataset	SST2	IMDB
samples	68221	50000
train	49118	0
validation	5458	0
test	13645	5848
class	2	2

Figure 1: Data sets statistics.

2.2 Baseline methods

We consider three baseline approaches for tackling the Out-of-Distribution (OOD) problem. Given an input x:

1. Maximum Soft-max Probability (MSP) (Hendrycks and Gimpel, 2016): $s_{MSP}(x) = 1 - max_{y \in Y} p_{Y|X}(y|x)$ where $p_{Y|X}(.|x)$ is the conditional soft-probability predicted by the model.

2. Energy-based score (E) (Liu et al., 2020): $s_E(x) = Tlog[\sum_{y \in Y} exp(\frac{g_y(x)}{T})]$ where $g_y(x)$ is the logit of the model corresponding to the class label y.

3. Mahalanobis (DM) (Podolskiy et al., 2021; Li et al., 2021; Zhou et al., 2021): $s_M(x) = -D_M(F(x), p_{F(X),\hat{y}})$ where D_M is the Mahalanobis score, $F(x)$ represents either the last layer or the probits of the encoder, and $p_{F(X),\hat{y}}$ is the training distribution of features with the same predicted class $\hat{y}$ as x.

For practical purposes, we set $T = 1$ for the temperature parameter in (E), and we implement $S = -s$ for (DM) and (E).

2.3 Pretrained model

We adapted a publicly available implementation[2] of BERT (Devlin et al., 2019) to our framework. The model is trained with a batch size of 16, weight decay set to 0.01, and learning rate set to 2×10^{-5} using the ADAMW optimizer (Kingma and Adam, 2014). The average test accuracy achieved (93%) is very close to the value reported in (Colombo et al., 2022a).

3 Results

The results from Figures 2 to 6 show that DM on the last layer performed the best for OOD detection, followed by MSP and E. However, logit aggregation worsened the performance of DM. All classifiers, except for DM on the last layer, correctly identified most in-distribution and out-of-distribution samples based on AUROC. However, they showed low precision rates based on AUPR, indicating that they missed some true positive OOD samples.

[2]https://github.com/CSCfi/machine-learning-scripts/blob/master/examples/pytorch-imdb-bert.py

It is concerning that the AUROC curve of DM on the logits is below the diagonal for high threshold values in OOD detection. This means that the classifier has a high rate of correctly identifying negative samples but has difficulty detecting positive samples, resulting in many false negatives.

Score	Input	AUROC	AUPR	FPR	ERR
MSP	softmax	80,9	57,4	50,2	27,7
E	logits	78,3	57,3	64,4	30
DM	logits	62,4	28,8	97,5	15,3
	last layer	86,0	64,5	62,2	11,6

Figure 2: OOD performance in %

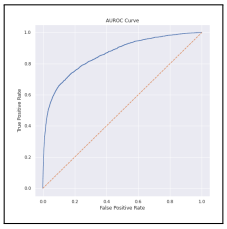

Figure 3: ROC curve-DM on last layer

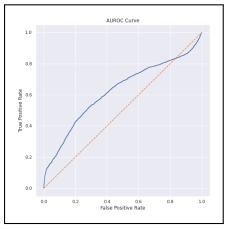

Figure 4: ROC curve-DM on logits

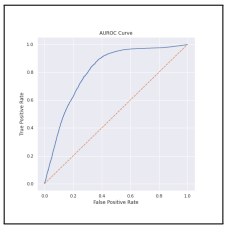

Figure 5: ROC curve-MSP

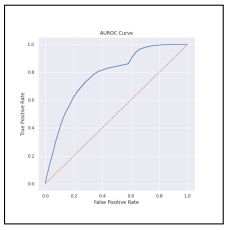

Figure 6: ROC curve-E

4 Discussion/Conclusion

The study examined the effectiveness of various out-of-distribution (OOD) detection methods using a single BERT model on the SST2 and IMDB datasets. The results showed that the deep features obtained from the DM method on the last layer outperformed other methods like MSP and E. These findings underscore the importance of using deep features to improve the performance of OOD detection.

However, the study also found that combining logits from multiple layers did not always lead to improved performance. Furthermore, most classifiers, except DM on the last layer, struggled to identify true positive OOD samples, indicating that further research is needed to improve the sensitivity of these methods by the use of robust measures (Staerman et al., 2021).

Despite the promising results, it is important to acknowledge some limitations of the study. For instance, the use of a single BERT model and limited datasets may restrict the generalizability of the findings to other NLP tasks and datasets. Additionally, the study was limited by the use of a single GPU, which prevented the exploration of more complex models or larger datasets, potentially hindering the discovery of better OOD detection methods.

In conclusion, while the study provided useful insights into the performance of different OOD detection methods using a BERT model, further research is needed to improve the sensitivity of these methods, increase the generalizability of the findings to other NLP tasks and datasets, and explore more complex models and larger datasets to advance OOD detection.

References

Eduardo Dadalto Câmara Gomes, Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. A functional perspective on multi-layer out-of-distribution detection.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition, 30(7):1145–1159*.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. *In Proceedings of the 23rd international conference on Machine learning, pages 233–240*.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642*.

Diederik P Kingma and Jimmy Ba. Adam. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Afshin Gholamy, Vladik Kreinovich, and Kosheleva. 2018. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *Departmental Technical Reports (CS). 1209.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *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 4171–4186*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *arXiv preprint arXiv:1610.02136*.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:1610.02136*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021a. Improving multimodal fusion via mutual dependency maximisation. *() EMNLP 2021*.

Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. 2021. k folden: k-fold ensemble for out-of-distribution detection. *arXiv preprint arXiv:2108.12731*.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *() ACL 2021*.

Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955.

Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. 2021. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv e-prints*, pages arXiv–2103.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. *arXiv preprint arXiv:2101.03778*.

Udit Arora, William. Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10687–10701*.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

Pierre Colombo, Eduardo D. C. Gomes, Guillaume Staerman, Noiry Nathan, and Pablo Piantanida. 2022a. Beyond mahalanobis-based scores for textual ood detection. *arXiv:2211.13527v1*.

Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *() ICML 2022*.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022b. Learning disentangled textual representations via statistical measures of similarity. *() ACL 2022*.

Marine Picot, Nathan Noiry, Pablo Piantanida, and Pierre Colombo. 2023a. Adversarial attack detection under realistic constraints.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023a. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*.

Marine Picot, Guillaume Staerman, Federica Granese, Nathan Noiry, Francisco Messina, Pablo Piantanida, and Pierre Colombo. 2023b. A simple unsupervised data depth-based method to detect adversarial images.

Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023b. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.