Detection of textual adversarial Attacks : Benchmark via Out-Of-Distribution examples identification

Nick Jofrein TEDONZE
ENSAE
nick.tedonze@ensae.fr

Boubacar SIDIBE
ENSAE
boubacar.sidibe@ensae.fr

Abstract

Word-level adversarial attacks have shown success in NLP models, drastically decreasing the performance of transformer-based models. Still, relatively few efforts have been made to detect adversarial examples, even if detecting adversarial examples may be crucial for automated NLP tasks especially for critical applications and models robustness. Pre-trained Transformers achieve high accuracy on in- distribution examples, and also in recent papers have shown to generalize better on out-of-distribution sample than previous models. In this work, we aim at detecting adversarial attacks in Natural Language Processing through Out-Of-Distribution (OOD) detection methods : Maximum Softmax Probability, DOCTOR detector and Mahalanobis distance-based score, on pre-trained Transformers such as BERT and Roberta. In the benchmark we provide, we generate 2 types of attacks on 4 datasets, and evalute the performance with AUROC and AUPR metrics. Our experimental results show that applying these simple out-of-distribution detection scores can provide acceptable performances for adversarial attacks detection. We provide code for our work, github.

1 Introduction

Deep learning is a powerful technology that has revolutionized various fields, including computer vision, natural language processing, and speech recognition. However, as deep learning models become more sophisticated and widely used, it has become increasingly important to ensure that these models are robust, fair, and secure (Colombo, 2021).

Fairness (Colombo et al., 2022b; Pichler et al., 2022), out-of-distribution (OOD) detection (Darrin et al., 2023a,b; Gomes et al.), and adversarial attack detection (Picot et al., 2023a,b) are three critical aspects of deep learning that require attention to ensure the reliability and safety of these systems. Fairness ensures that deep learning models do not discriminate against individuals or groups based on their characteristics, such as race, gender, or religion. OOD detection is important to identify when a model is presented with inputs that are outside the range of what it was trained on. Adversarial attack detection ensures that models are robust against attempts to manipulate or exploit them by attackers. In this work, we choose to focus on safety against a malicious attack in the scope of NLP models.

In natural language processing, adversarial examples refer to seemingly innocent texts that alter the model prediction to a desired output, yet remain imperceptible to humans. Generated adversarial examples maintain the semantic similarity well and in recent years word-level attacks have shown success in NLP models, drastically decreasing the performance of transformer-based models in sentence classification tasks (Li et al., 2020; Ren et al., 2019). In the NLP research community, some progress has been made for defending against the attacks (Keller et al., 2021; Zhou et al., 2021), but only few efforts have been made in techniques for their detection.

Our contribution try to remediate to that, following (Yoo et al., 2022a) benchmark on four popular attacks on 4 text classification datasets and four models, plus their Robust density estimation detection method, we use pre-trained transformer-based models that have shown success in in-sample prediction and to generalize better to OOD samples that previous models, by providing a benchmark on 2 highly effective word-level attacks on 4 popular sequence classification datasets using Maximum Softmax probability (Hendrycks and Gimpel, 2016), DOCTOR discriminator (Granese et al., 2021) and Maha-

lanobis distance-based detection methods (Podolskiy et al., 2021).

2 Problem Framing

Given an input space $\mathcal{X}$ and a label space $\mathcal{Y}$, a predictive model $F : \mathcal{X} \rightarrow \mathcal{Y}$, and an oracle model $F^* : \mathcal{X} \rightarrow \mathcal{Y}$, a successful adversarial example x_{adv} of an input $x \in \mathcal{X}$ satisfies the following :

$$F^*(x) = F(x) \neq F(x_{adv}),$$
$$C_i(x, x_{adv}) = 1 \text{ for } i \in \{1, ..., n\}$$

where C_i is an indicator function for the i-th constraint between the perturbed text and the original text, which is 1 when the two texts are indistinguishable with respect to the constraints varying from attack algorithms.
Let $D_n := \{(x_i, y_i)\}_{i=1}^n$ be the collection of text inputs. Assuming that adversarial samples come from a different distribution than the one that generates D_n, we construct an out-of- domain detector that takes an unseen new input w and determines whether w comes from the same distribution that generates D_n. Mathematically, the OOD detector can be written as a function :

$$g(w, \epsilon) = \begin{cases} True & \text{if } S(w) \leq \epsilon \\ False & \text{if } S(w) > \epsilon \end{cases}$$

where $S()$ denotes the anomaly score function, and ϵ is a chosen threshold to ensure that the true positive (in-domain) rate is at a certain level (e.g., 95%). The OOD detection problem boils down to designing $S()$ such that it assigns in-domain inputs lower scores than out-of-domain inputs.

3 Experiments Protocol

For evaluating the detection scores, we generate attacks (TF & PWWS) on 4 datasets with respect to 2 transformer-based models, BERT and Roberta. As training these models is highly computationally expensive, and our goal is to work on pre-trained transformers, we generate the attacks using TextAttack library. This library allows generating adversarial attacks on various dataset on transformer-based model, including the ones we are considering. Once the adversarial attacks are generated for each model-dataset-attack combination, we train our implemented model for the classification task, after we test it on a sample combining adversarial examples and simple test-examples, and then compute our detection scores. Finally, we evaluate the methods with AUROC and AUPR metrics.

3.1 Datasets

Following benchmarks based on (Yoo et al., 2022b) and (Hendrycks et al., 2020), we choose four sentence classification datasets :

- **IMDB** : a dataset of polarized movie reviews with 25000 training and 25000 test reviews (Maas et al., 2011) and **SST-2** : which contains pithy expert movie reviews (Socher et al., 2013). Models predict a movie review's binary sentiment, and we report accuracy.

- **YELP review dataset** : contains restaurant reviews with detailed metadata (e.g., user ID, restaurant name). As in (Hendrycks et al., 2020) we carve out four groups from the dataset based on food type: American, Chinese, Italian, and Japanese. Models predict a restaurant review's binary sentiment, and we report accuracy.

- **AG-News** : sub-dataset of AG's corpus of news articles constructed by assembling titles and description fields of articles from the 4 largest classes ("World", "Sports", "Business", "Sci/Tech") (Zhang et al., 2015).

3.2 Attacks

In this work, we consider two widely known attacks, Textfooler (Jin et al., 2019) and Probability Weighted Word Saliency (Ren et al., 2019). On a sample of each of the dataset, attacks are generated for BERT and Roberta models, and label 1 for successful attacks and 0 else (TF and PWWS nearly achieve 100% success rate). Theses adversarial examples were generated beforehand using TextAttack library (Morris et al., 2020).

3.3 Detection methods

3.3.1 Related works

First let us have a word for some of the recent detection methods developed recent papers are the following. Frequency-guided word substitutions (FGWS), a simple algorithm exploiting the frequency properties of adversarial word substitutions for the detection of adversarial examples (Mozes et al., 2021).
Robust Density Estimation (RDE), a detection method that utilizes density estimation (Yoo et al., 2022a). The principle is to fit a parametric density estimation model in the feature space, for modeling the probability distribution of raw

Dataset	Topic	Task	Classes	# of test samples
IMDB	movie review	sentiment classification	2	25K / 10K
YELP	restaurant review	sentiment classification	2	38K / 5K
AG-News	news headline	topic classification	4	7.6K / 7.6K
SST-2	movie review	sentiment classification	2	2.7K / 2.7K

Table 1: Summary of the benchmark dataset.

texts, and then evaluate the likelihood of a test sample to assess how far/or not it deviates from the distribution of the training set.
Maximum Likelihood Estimator denoted as MLE, which is a out-of-distribution detector from the image domain. Similar to RDE, it fits a Gaussian model using the maximum likelihood estimation (MLE) then trains a logistic regressor using the likelihood scores (Lee et al., 2018).

In our benchmark we will use the following detection scores/methods.

3.3.2 Maximum Softmax probability

Maximum Softmax probability (Hendrycks and Gimpel, 2016) consists in retrieving the maximum/predicted class probability from a softmax distribution, and thereby detect whether an example is erroneously classified or out-of-distribution. Specifically, it assigns an example x the anomaly score $-\max p(y|x)$. Specifically for pre-trained transformers, given an text input x it consists in computing :

$$\mathcal{I}(x) := -\max_{i\in[K]} \frac{\exp(h_i(x)/T)}{\sum_{j=1}^{K}\exp(h_j(x)/T)}$$

where $h_i(x)$ is the output logits layer of the multi-class classifier, and T is the temperature that is selected such that the true positive rate is at a given rate (e.g., 95% in (Liang et al., 2017)).

3.3.3 DOCTOR discriminator

DOCTOR is a out-of-distribution detector based on the softmax probability of an input x given by the last layer a DNN, a method that aims to identify whether the prediction of a deep neural networks classifier should (or should not) be trusted (Granese et al., 2021), it does not require prior information about the underlying dataset. Given x it can be defined as :

$$D_\alpha(x) = \sum_{y\in\mathcal{Y}} softmax^2(x)_y$$

3.3.4 Mahalanobis-based score

Mahalanobis's distance can be seen as a depth measure (Staerman et al., 2021), notion of statistical depth (). It measures the distance between an element in $\mathbb{R}$ and a probability distribution having finite expectation and invertible covariance matrix. Precisely, the Mahalanobis depth function is defined as:

$$M(x) = \max_{y\in\mathcal{Y}} -(f(x)-\mu_c)^T\Sigma^{-1}(f(x)-\mu_c)$$

where $f(x)$ is the logits vector of the entry x, μ_c is the class mean and Σ is the covariance matrix. Variety of OOD detectors can be developed based on Mahalanobis distance, as for instance *TRUSTED* detector which has outperformed previous approaches (Colombo et al., 2022a).

3.4 Performance metrics

There exist several ways to measure the effectiveness of an OOD method, we will apply them to assess the performance for adversarial attacks detection.
Area Under the Receiver Operating Characteristic curve (AUROC) (Bradley, 1997) which is a threshold-independent performance evaluation. It is the area under the ROC curve which plots the false positive rate against the true positive rate. AUROC can be interpreted as the probability that a positive example has a greater detector score than a negative example, a random positive example detector corresponds to a 50% AUROC, and a "perfect" classifier corresponds to 100%.
Area Under the Precision-Recall curve (AUPR), which plots the precision and recall against each other. The AUPR is known to be more relevant to unbalanced situations (Davis and Goadrich, 2006).

4 Experimental results

In this section, we present the result we obtain on the 4 datasets using 3 detection methods on 2 attacks.
Before diving into our numerical results, let us present an empirical analysis of the behavior of MSP, DOCTOR, and MHLNB when faced with the task of choosing whether to accept or reject the prediction of a given classifier for a certain sample.

In Figure 1, we propose a graphical interpretation of the discrimination performance, we represent the distribution of the detectors' scores according to their true labels in blue and in orange, respectively. We see that the discrimination performance of scores is quite mitigated, especially for Mahalanobis distance-based score.

Table 2 and 3 demonstrate the results for two model-attack combination on all datasets. Maximum softmax probability, DOCTOR score and Mahalanobis distance-based score are able to distinguish between adversarial samples and normal samples in expectation as shown by the higher-than-random metrics.

Table 2: Results table for Textfooler attack - RoBERTa.

Dataset	**Method**	**AUROC %**	**AUPR %**
IMDB	MSP	75	74.2
	DOCTOR	74.3	74.4
	MHLNB	62.4	45
SST-2	MSP	82.6	80.5
	DOCTOR	81.9	80.8
	MHLNB	82.4	80.7
YELP	MSP	78.8	77.0
	DOCTOR	78.9	76.6
	MHLNB	80.9	79.6
AG-NEWS	MSP	64.4	62.0
	DOCTOR	64.7	63.7
	MHLNB	70.4	69.9

Table 3: Result table for PWWS attack - BERT.

Dataset	**Method**	**AUROC %**	**AUPR %**
IMDB	MSP	76.2	69.5
	DOCTOR	74.4	70.5
	MHLNB	70.3	79.8
SST-2	MSP	80.2	69.3
	DOCTOR	81.8	75.8
	MHLNB	80.1	78.1
YELP	MSP	75.6	71.2
	DOCTOR	79.5	78.7
	MHLNB	80.2	81.5
AG-NEWS	MSP	69.7	64.5
	DOCTOR	70.1	67.6
	MHLNB	70.8	68.0

We notice that in general, score based on information available from the training set (MHLNB) achieve stronger results than those relying on output of softmax scores (MSP, DOCTOR). This observation is comforted by previous research (Podolskiy et al., 2021).

5 Discussion/Conclusion

In conclusion, we have presented a set of simple yet effective methods for detecting adversarial examples in natural language processing (NLP) tasks, building on the foundation of out-of-distribution detection. Our experimental results demonstrate that using these out-of-distribution detection scores as adversarial attack detectors yields acceptable performance on average. However, these methods may still not be reliable enough for use in critical applications.

Future research directions could focus on improving the performance of the proposed methods, by exploring more sophisticated out-of-distribution detection techniques or incorporating additional features such as semantic information or context awareness. Another direction could be to investigate the transferability of the proposed methods across different NLP tasks and models, or to explore their applicability to other domains such as computer vision or speech recognition. Additionally, it would be worthwhile to investigate the potential of combining multiple detection methods to achieve higher detection accuracy and robustness. Overall, there is ample room for future research to advance the state-of-the-art in adversarial example detection and enhance the security and reliability of NLP systems.

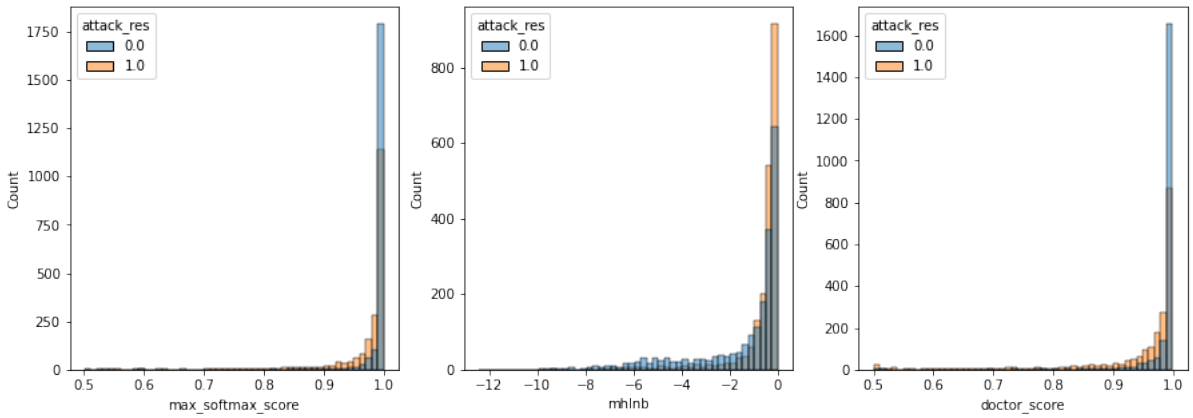

Figure 1: DOCTOR, ODIN, SR and MHLNB to split data samples in IMDB for TF attack - BERT

.

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 Recognit.*, 30:1145–1159.

Jesse Davis and Mark H. Goadrich. 2006. The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*.

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

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, Seattle, Washington, USA. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *ArXiv*, abs/1509.01626.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv: Learning*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Red Hook, NY, USA. Curran Associates Inc.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Association for Computational Linguistics*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Conference on Empirical Methods in Natural Language Processing*.

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online. Association for Computational Linguistics.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *ACL*.

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

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 171–186, Online. Association for Computational Linguistics.

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

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.

A. V. Podolskiy, Dmitry Lipin, A. Bout, E. Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *AAAI Conference on Artificial Intelligence*.

Pierre Colombo, Eduardo Gomes, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022a. Beyond mahalanobis-based scores for textual ood detection. *ArXiv*, abs/2211.13527.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022a. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022b. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *arXiv*.

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*.