Social Computing Group,
Department of Informatics
Technical University of Munich

Detecting Word-Level Adversarial Text Attacks via SHapley Additive exPlanations

Lukas Huber*, Marc Alexander Kühn*, Edoardo Mosca*, Georg Groh

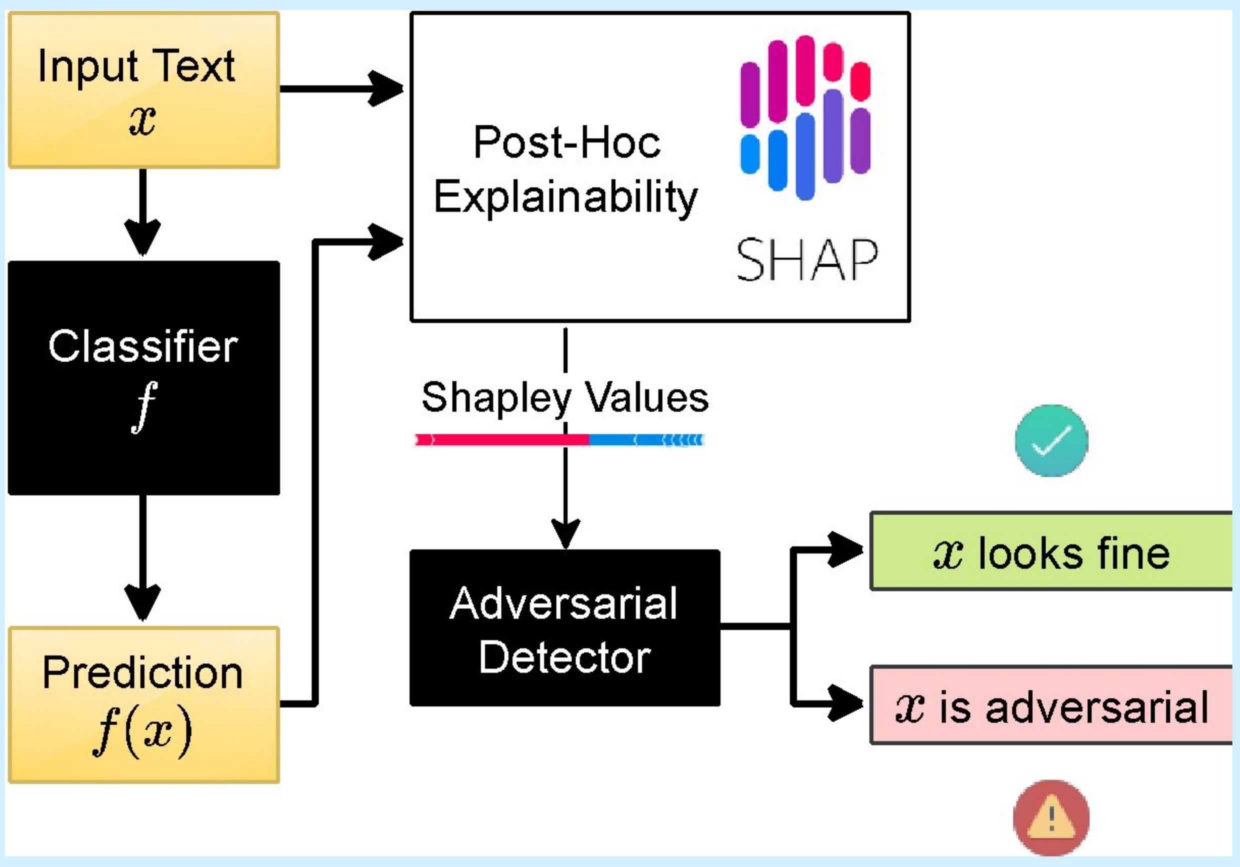
Motivation

- State-of-the-art machine learning models are prone to adversarial attacks: Maliciously crafted model inputs to fool the prediction
- Research in NLP still lacks techniques to make models resilient against those attacks
- We adapt a method from computer vision to detect word-level attacks leveraging SHAP

SHAP

- Based on the game-theoretical concept of Shapley values
- Allows to score the contribution of every word towards the overall prediction
- Adversarial attacks change characters/words to change the prediction. The modified tokens have large influence on the predicted class.
- The SHAP values for a whole sentence is called a signature

Method

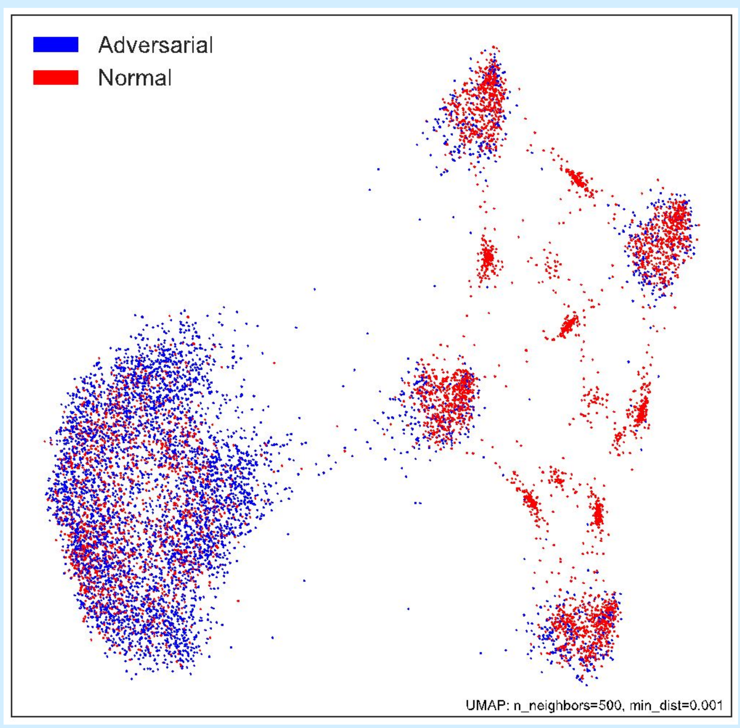


1) Compute the SHAP signatures for the unaltered samples
2) Generate adversarial samples and compute their signature
3) Train adversarial detector signatures to differentiate between adversarial and normal inputs.
4) For unseen samples, calculate the signature of the input and feed it into the model.

Results

	Method	AG_News	IMDb	SST-2	Yelp Polarity	Metric
Our	Neural Network	0.90 / 0.90	**0.96 / 0.96**	0.75 / 0.75	**0.94 / 0.94**	F1 score / Accuracy
	Random Forest	**0.91 / 0.91**	0.87 / 0.87	**0.77 / 0.77**	0.84 / 0.84	F1 score / Accuracy
	SVM	0.90 / 0.90	0.90 / 0.90	0.74 / 0.74	0.89 / 0.89	F1 score / Accuracy
SotA Detector	FGWS [1]	-	0.77	0.63	-	F1 score
Other Defenses	DNE [2]	**0.91**	0.82	-	-	Accuracy
	SEM [3]	0.76	0.85	-	-	Accuracy
	ASCC [4]	-	0.77	-	-	Accuracy

Base-Model	IMDb (Test)	SST-2 (Test)
IMDb	-	0.56
SST-2	0.42	-
Yelp Polarity	**0.71**	**0.66**



- We outperform the state-of-the-art detector and all other defenses
- Our detector is in some cases transferable to other datasets
- SHAP signatures of most adversarial samples collapse into a single cluster

Conclusion

- Leveraging SHAP explanations for detecting adversarial examples works well for NLP
- Model explanations explicitly encode information to separate attacks from their conterpart
- Regarding transferability, our results are promising but not sufficient
- Future research should focus on performance evaluation against multiple types of attacks and models plus generalization across multiple datasets

SCAN ME

*** Authors with equal contribution.**

[1] Maximilian Mozes, Pontus Stenetorp, Bennett Klein- berg, 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 Compu- tational Linguistics: Main Volume, pages 171–186, Online. Association for Computational Linguistics.
[2] Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-wei Chang, and Xuanjing Huang. 2020. Defense against adversarial attacks in nlp via dirichlet neighborhood ensemble. arXiv preprint arXiv:2006.11627.
[3] Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural language adversarial attacks and defenses in word level. arXiv preprint arXiv:1909.06723.
[4] Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural lan- guage word substitutions. In 9th International Con- ference on Learning Representations (ICLR).