BOHR

BOHR International Journal of Computer Science
2022, Vol. 1, No. 1, pp. 26–31
DOI: 10.54646/bijcs.2022.04
www.bohrpub.com

METHODS

White-box attacks on hate-speech BERT classifiers in German with explicit and implicit character-level defense

Shahrukh Khan[*]**, Mahnoor Shahid and Navdeeppal Singh**

[*]**Correspondence:**
Shahrukh Khan,
shkh00001@stud.uni-saarland.de

Received: 24 February 2022; **Accepted:** 02 March 2022; **Published:** 15 March 2022

Attention-based transformer models have achieved state-of-the-art results in natural language processing (NLP). However, recent work shows that the underlying attention mechanism can be exploited by adversaries to craft malicious inputs designed to induce spurious outputs, thereby harming model performance and trustworthiness. Unlike in the vision domain, the literature examining neural networks under adversarial conditions in the NLP domain is limited and most of it focuses mainly on the English language. In this article, we first analyze the adversarial robustness of Bidirectional Encoder Representations from Transformers (BERT) models for German data sets. Second, we introduce two novel NLP attacks: a character-level and a word-level attacks, both of which utilize attention scores to calculate where to inject character-level and word-level noise, respectively. Finally, we present two defense strategies against the attacks above. The first implicit character-level defense is a variant of adversarial training, which trains a new classifier capable of abstaining/rejecting certain (ideally adversarial) inputs. The other explicit character-level defense learns a latent representation of the complete training data vocabulary and then maps all tokens of an input example to the same latent space, enabling the replacement of all out-of-vocabulary tokens with the most similar in-vocabulary tokens based on the cosine similarity metric.

Keywords:

Introduction

Natural language processing (NLP) has achieved tremendous progress in surpassing human-level baselines in a plethora of language tasks with the help of attention-based neural architectures (1). However, recent studies (2–4) show that such neural models trained via transfer learning are susceptible to adversarial noise. However, this also presents new challenges against adversaries which pose a realistic threat to machine learning system's utility if present. This is because attention attributions can be potentially be exploited by an adversary to craft attacks that require least perturbation budget and compute to carry out a successful attack on the victim neural network model. Moreover, to the best of our knowledge, most work concentrates on English language corpora.

Adversarial attacks on machine learning models are possible to defend against while also minimizing risks to degradation of model's utility and performance. Two novel defense strategies are proposed: implicit and explicit character-level defenses. Implicit character-level defense introduces a variant of adversarial training where the adversarial text sequences are generated via white-box character-level attack and are mapped to a new abstain class and then the model is retrained. Whereas explicit character-level defense performs adversarial pre-processing of each text sequence prior to inference to eliminate adversarial signals, hence resulting in transformation of adversarial input to benign.

Literature survey

Hsieh et al. (2) proposed using self-attention scores for computing token importances in order to rank potential candidate tokens for perturbation. However, one potential

TABLE 1 | Data set statistics.

Data set	Train	Validation	Test
HASOC 2019	3054	765	850
GermEval 2021	2594	650	944

TABLE 2 | Undefended models.

Dataset	Accuracy(%)
HASOC 2019	84
GermEval 2021	69

TABLE 3 | Character-level attack on defended models.

Data set	Defense	Attack success rate(%)
HASOC 2019	Explicit character level	9.5
GermEval 2021		**5.3**
HASOC 2019	Implicit abstain based	**1**
GermEval 2021		11.1

shortcoming of their idea is that they replace the potential token candidate with random tokens from vocabulary, which may result in changing the semantic meaning of perturbed sample. Garg et al. (3) proposed BERT-based Adversarial Examples for Text Classification in which they employ Mask Language Modeling (MLM) for generating potential word replacements in a black-box setting. Finally, Pruthi et al. (4) showed susceptibility of BERT (5) based models to character-level misspellings also in a black-box setting. In our study, we employ both character-level and word-level attacks in a white-box setting.

Problem statement

To use attention mechanism in transfer learning setting to craft word and character-level adversarial attacks on neural networks. Also, evaluate and compare the robustness of two novel character-level adversarial defenses.

Experimental setting

Undefended models

Data sets

We present our work based on HASOC 2019 (German Language) (6) and GermEval 2021 (7). Both of the sub-tasks are binary classification tasks where the positive labels correspond to hate-speech and negative labels correspond to non-hate-speech examples (**Table 1**).

Training

For training, the undefended models, we fine-tune GBERT (8) language model for German language which employs training strategies, namely, Whole Word Masking (WWM) and evaluation-driven training and currently achieves SoTA performance for document classification task for German language. We obtain the following accuracy scores for each data set, respectively (**Table 2**).

Attacks

Baseline word-level white-box attack

The baseline word-level attack is composed by enhancing token candidate proposed by Hsieh et al. (2) which prominently replaces tokens sorted in order of their attention scores with random tokens from vocabulary, which may lead to perturbed sequence being semantically dissimilar to the source sequence. In the baseline attack, we address this potential shortcoming by using a language model MLM to generate potential candidate for each token ranked in the order of attention scores. Furthermore, instead of just performing the replacement operation, we employ the perturbation scheme proposed by Garg et al. (3) and insert generated tokens to the left/right of the target token where the candidate tokens are generated via MLM.

Word-level White-box attack

The main motivation behind this attack is based on the fact using only language models to ensure semantic correctness in the adversarial sequences is not enough. Since it highly depends on the vocabulary of the pretrained language model. We improve the baseline attack for the preserving more semantic and syntactic correctness of the source sequence by introducing further constraints on the generated sequence by the baseline attack. First, we compute the document-level embeddings for both perturbed and source sequences and then compute cosine similarity with a minimum acceptance threshold of 0.9363 as originally suggested by Jin et al. (9), since Garg et al. (3) developed their work using the same threshold value. Finally, we further add another constraint that Part of Speech (POS) tag of both candidate and target tokens should be same.

Character-level White-box attack

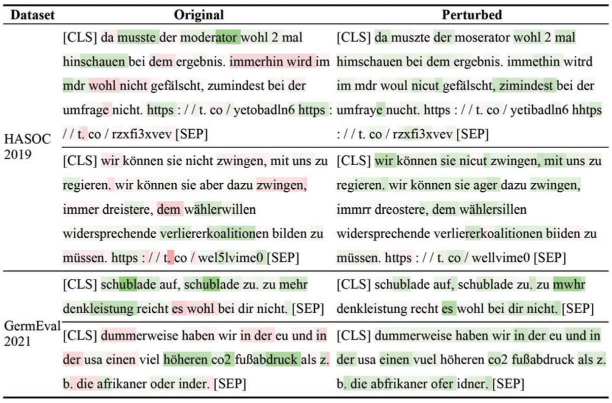
In this white-box character-level attack, attention scores are obtained in order to get the word importance, similar to earlier white-box word-level attacks. Then, by ordering the word importance in the order of higher to lower, we employ the character perturbation scheme employed by Pruthi et al. (4) since they evaluated this in the black-box setting only. In our study, we perform character-level perturbation within a target token by token modification of character (e.g., swap, insert, and delete) applied to cause perturbations such adversarial examples are utilized to maximize the change in model's original prediction confidence with limited numbers of modifications. However, these modifications prove to be significantly effective, as outlined in section "Results."

Defenses

Abstain-based training

In several past evaluations and benchmarks of defenses against adversarial examples (10–15), adversarial training(16) has been found tobeoneofthebestways of conferring robustness. However, it is computationally expensive due to the need of creating adversarial examples during training. Thus, we chose to employ a detection-based defense, which we call abstain-based training. Although detection-based defenses are known to be not as effective as adversarial training (11, 15), we still believe our method will deliver insights into the capability of BERT models in recognizing adversarial examples similar to adversarial training due the way it works. In contrast to other detection-based defenses in the literature (17–21), the approach is much simpler. It works as follows.

Let C be the trained undefended classifier. We create a new (untrained) classifier *C* from C by extending the number of classes it is able to predict by one. The new class is labeled "ABSTAIN," representing that the classifier abstains from making a prediction. Using C we create the adversarial examples. We mix these with the normal examples from the data set (of C), where the adversarial examples have the abstain label, to create a new data set. We then simply train on this data set. We applied this defense strategy on the models from Section "Training" and present the results in **Table 3**. We also show the classification attributions in **Figure 1** to try to interpret the models' behavior.



Dataset	Original	Perturbed
HASOC 2019	[CLS] da musste der moderator wohl 2 mal hinschauen bei dem ergebnis. immerhin wird im mdr wohl nicht gefälscht, zumindest bei der umfrage nicht. https : / / t. co / yetobadln6 https : / / t. co / rzxfi3xvev [SEP]	[CLS] da muszte der moserator wohl 2 mal himschauen bei dem ergebnis. immethin witrd im mdr woul nicut gefälscht, zimindest bei der umfraye nucht. https : / / t. co / yetibadln6 hhtps : / / t. co / rzxfi3xvev [SEP]
	[CLS] wir können sie nicht zwingen, mit uns zu regieren. wir können sie aber dazu zwingen, immer dreistere, dem wählerwillen widersprechende verliererkoalitionen bilden zu müssen. https : / / t. co / wel5lvime0 [SEP]	[CLS] wir können sie nicut zwingen, mit uns zu regieren. wir können sie ager dazu zwingen, immrr dreostere, dem wählersillen widersprechende verliererkoalitionen biiden zu müssen. https : / / t. co / wellvime0 [SEP]
GermEval 2021	[CLS] schublade auf, schublade zu. zu mehr denkleistung reicht es wohl bei dir nicht. [SEP]	[CLS] schublade auf, schublade zu. zu mwhr denkleistung recht es wohl bei dir nicht. [SEP]
	[CLS] dummerweise haben wir in der eu und in der usa einen viel höheren co2 fußabdruck als z. b. die afrikaner oder inder. [SEP]	[CLS] dummerweise haben wir in der eu und in der usa einen vuel höheren co2 fußabdruck als z. b. die abfrikaner ofer idner. [SEP]

FIGURE 1 | Visualization of the classification attributions of the abstain-based trained models, which correctly classify the examples. The perturbed examples shown above fool the normally trained models. We observe that the attributions are much more spread out when models encounters a perturbed example. (Words were split by the tokenizer, thus a single word can have different sub-attributions).

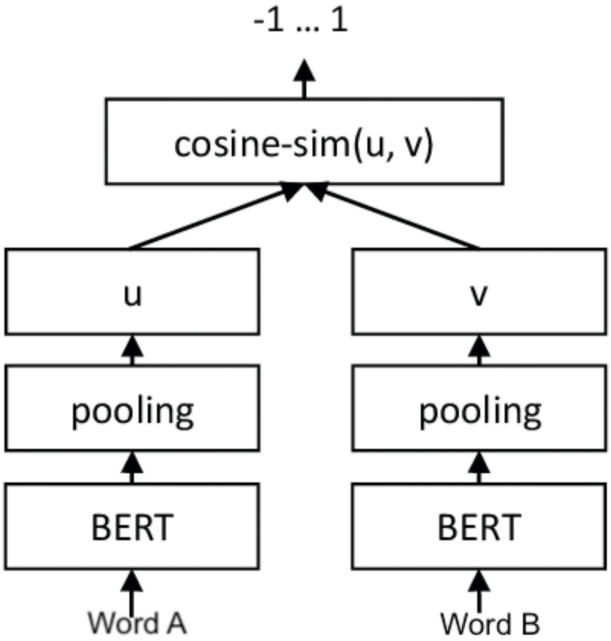


FIGURE 2 | Sentence-BERT for character level similarity.

TABLE 4 | Attacks result on undefended models.

Data set	Attack	Successrate(%)
HASOC 2019	Baseline	8.49
GermEval 2021		60.3
HASOC 2019	Word-level	4.03
GermEval 2021		49.8
HASOC 2019	Character-level	**73.1**
GermEval 2021		**93.5**

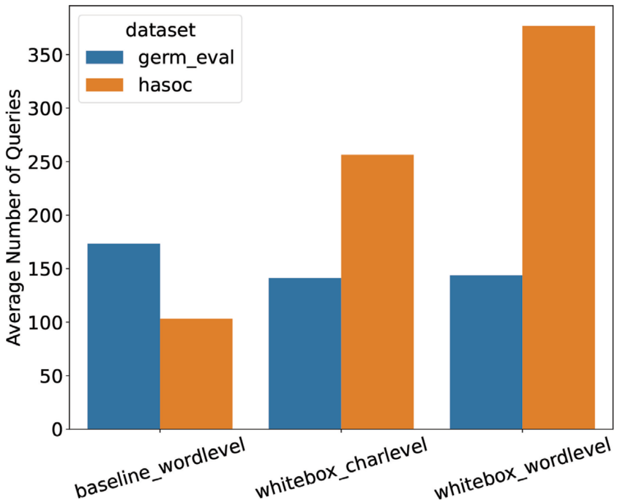


FIGURE 3 | Average number of queries per successful attack.

Explicit character-level defense

Abstain-based training defense achieves high success in defending against the adversarial character-level perturbed inputs. However, this results in degraded system utility since the model does not make any useful prediction when the input is perturbed at character level. To overcome this drawback, we propose the explicit characterlevel defense, which is an unsupervised approach which makes an assumption that

$\forall\, t \in T_{input} : t \in V_{train.}$

Here, V_{train} is the set of all tokens present in the training set. However, replacing this set with set of words in the given language, i.e., set of all words in German language etc., would result in better results. T_{input} refers to set of tokens present in the input sequence and we assume the worstcase,which means *Tinput* is perturbed with character-level noise.

In this defense method, we first re-purpose the Sentence-BERT (22) architecture, which originally trained sentence pairs to compute semantic vector representations and achieved SoTA results on multiple information retrieval data sets. However, we change input to character level by inputting word pairs to the network. Concretely, we labeled the Birkbeck spelling error corpus (23) has word pairs with one correct and the other misspelled word and we label each pair based on the Levenshtein distance between each pair. The schematics of our neural approach are given in **Figure 2**.

The main idea behind using the neural approach is to project similarly spelled words close to each other in the vector space. Algorithm 1 outlines main idea of our approach for explicit character-level defense.

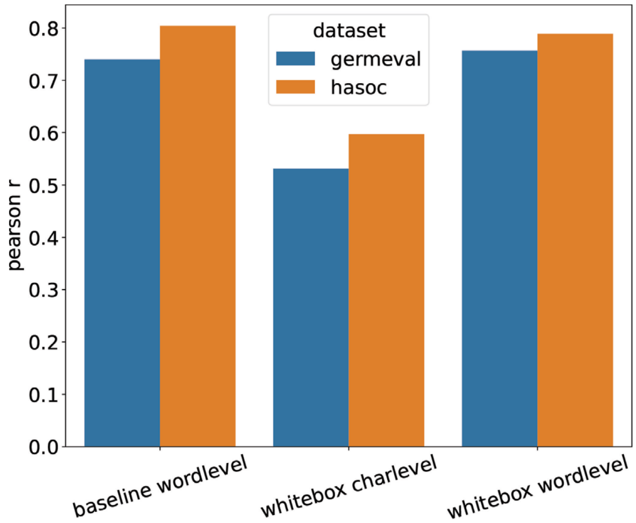


FIGURE 4 | Pearson correlation between original text length and number of queries for attack success.

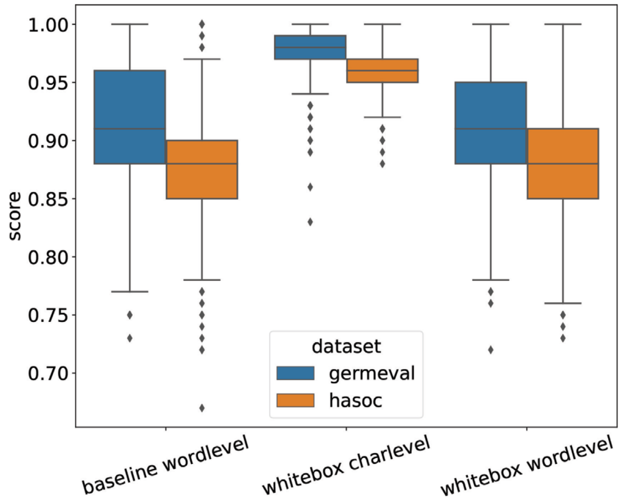


FIGURE 5 | Levenshtein distance-based similarity between original and perturbed sequences.

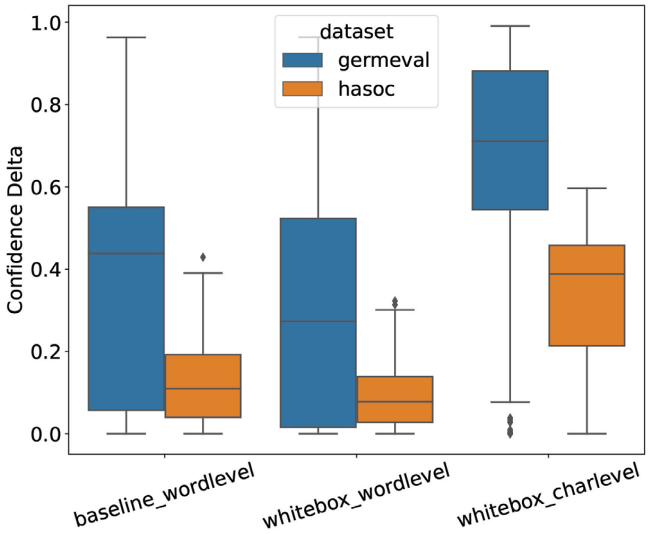


FIGURE 6 | Confidence Delta between original and perturbed sequences caused by each attack.

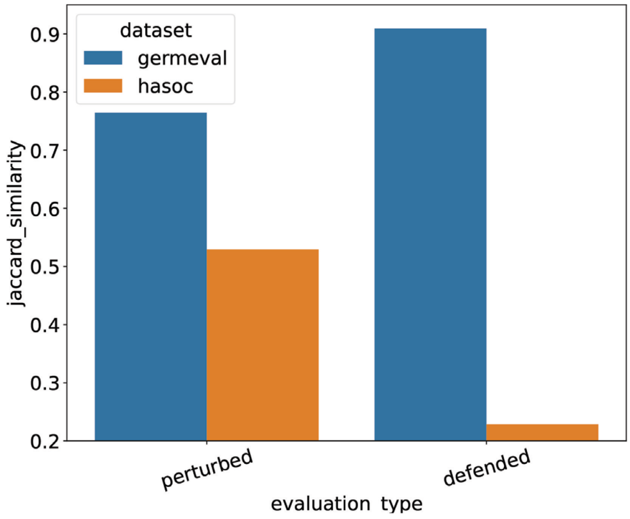


FIGURE 7 | Jaccard similarity between original and perturbed text vs. the original and defended text.

```
begin:
$V_{train}$ ti... $t_m$ > Set of tokens in vocabulary
$E_v$ el... em > Embeddings of vocabulary
Tinput ti... tj > Set of tokens in input
for k 1 to j do
ek vi... $v_n$ > Get embedding of input token k
scores cos($E_v$, ek)
 >  Cosine similarity with vocabulary embeddings
ifmax scores > 0.7 and max scores < 1.0
then $vocab_{in}d_{ex}$ arg max scores;
$T_{input}[k]$ $V_{train}$ $[vocab_{index}]$
end for
```

Algorithm 1 | Explicit character-level defense.

Results

Attack results

Table 4 shows the character-level attacks to be most effective on both models.

Figure 3 illustrates how the number of queries required per sample for a successful attack depends on the data set and attack type, we further show in Figure 4 that both word-level attacks require more queries for a longer sequence as compared to character-level attack, which is slightly agnostic to the sequence length. Figure 5 shows that the character-level attack requires minimal amount of perturbation since the changes are at word level; moreover from Figure 6, it can be concluded that character-level attack also makes the highest difference in model prediction confidence in case of a successful attack.

Defense results

Conclusion

We show that self-attentive models are more susceptible to character-level adversarial attacks than word-level attacks on text classification NLP task. We provide two potential ways to defend against character-level attacks. Future work can be done to enhance the explicit character-level defense using supervised sequence-to-sequence neural approaches, as can be seen in Figure 7 where current approach enhances the jaccard similarity of defended sequences with original sequences when compared to jaccard similarity between original sequence and perturbed sequence in case of GermEval 2021. However, for HASOC 2019 data set because of abundance of out-of-vocabulary tokens in the unseen test set, the defense degrades the quality of defended sequences. However, even then the defense proves to be quiet robust against character-level adversarial examples, as can be seen in Table 3.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. *ArXiv [Preprint]* (2017):doi: 10.48550/arXiv.1706.03762
2. Hsieh Y, Cheng M, Juan D, Wei W, Hsu W, Hsieh C. On the robustness of self-attentive models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics (2019). p. 1520–9.
3. Garg S, Ramakrishnan G. BAE: BERT-based adversarial examples for text classification. *ArXiv [Preprint]* (2020):doi: 10.48550/arXiv.2004.01970
4. Pruthi D, Dhingra B, Lipton Z. Combating adversarial misspellings with robust word recognition. *ArXiv [Preprint]* (2019):doi: 10.48550/arXiv.1905.11268
5. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv [Preprint]* (2018):doi: 10.48550/arXiv.1810.04805
6. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, et al. Overview of the hasoc track at fire 2019: hate speech and offensive content identification in indo-european languages. *Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE '19*. New York, NY: Association for Computing Machinery (2019). p. 14–7.
7. Risch J, Stoll A, Wilms L, Wiegand M. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. Duesseldorf: Association for Computational Linguistics (2021). p. 1–12.
8. Chan B, Schweter S, Meiller T. German's next language model. *ArXiv [Preprint]* (2020):doi: 10.48550/arXiv.2010.10906
9. Jin D, Jin Z, Zhou J, Szolovits P. Is BERT really robust? Natural language attack on text classification and entailment. *ArXiv [Preprint]* (2019):doi: 10.48550/arXiv.1907.11932
10. Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. *Proceedings of the International conference on machine learning. PMLR* (2018). p. 274–83.
11. Carlini N, Wagner D. Adversarial examples are not easily detected: bypassing ten detection methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. New York, NY: (2017).
12. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*. San Jose, CA: (2017). p. 39–57.
13. Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research*. PMLR (2020). p. 2206–16.
14. Croce F, Andriushchenko M, Sehwag V, Flammarion N, Chiang M, Mittal P, et al. Robustbench: a standardized adversarial robustness benchmark. *ArXiv [Preprint]* (2020):doi: 10.48550/arXiv.2010.09670
15. Bryniarski O, Hingun N, Pachuca P, Wang V, Carlini N. Evading adversarial example detection defenses with orthogonal projected gradient descent. *ArXiv [Preprint]* (2021):doi: 10.48550/arXiv.2106.15023

16. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. *ArXiv [Preprint]* (2018):doi: 10.48550/arXiv.1706.06083

17. Grosse K, Manoharan P, Papernot N, Backes M, McDaniel P. On the (statistical) detection of adversarial examples. *ArXiv [Preprint]* (2017):doi: 10.48550/arXiv.1702.06280

18. Gong Z, Wang W, Ku W. Adversarial and clean data are not twins. *ArXiv [Preprint]* (2017):doi: 10.48550/arXiv.1704.04960

19. Bendale A, Boult T. Towards open set deep networks. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Las Vegas, NV: IEEE Computer Society (2016). p. 1563–72.

20. Sotgiu A, Demontis A, Melis M, Biggio B, Fumera G, Feng X, et al. Deep neural rejection against adversarial examples. *EURASIP J Inform Secur.* (2020) 2020:1–10.

21. Metzen J, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. Toulon: (2017).

22. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics (2019). 11 p.

23. Mitton R. *Birkbeck Spelling Error Corpus*. Oxford: University of Oxford (1980).