BOHR International Journal of Intelligent Instrumentation and Computing
2021, Vol. 1, No. 1, pp. 24–29
https://doi.org/10.54646/bijiiac.004
www.bohrpub.com

White-Box Attacks on Hate-speech BERT Classifiers in German with Explicit and Implicit Character Level Defense

Shahrukh Khan, Mahnoor Shahid and Navdeeppal Singh

shkh00001@stud.uni-saarland.de; mash00001@stud.uni-saarland.de; s8nlsing@stud.uni-saarland.de

Abstract. 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 paper, we first analyze the adversarial robustness of Bidirectional Encoder Representations from Transformers (BERT) models for German datasets. Second, we introduce two novel NLP attacks. Namely, 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:

1 Introduction

Natural language processing has achieved tremendous progress in surpassing human-level baselines in a plethora of language tasks with the help of attention based neural architectures [2]. However, recent studies [3, 4, 5] 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. As 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 Implicit and Explicit Character-Level defenses are proposed. 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 preprocessing of each text sequence prior to inference to eliminate adversarial signals hence results in transformation of adversarial input to benign.

2 Literature Survey

Hsieh et al. [3] proposed using self attention scores for computing token importances in order to rank potential candidate tokens for perturbation. However, one potential shortcoming of their idea is 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. [4] proposed BERT-based Adversarial Examples for Text Classification in which they employ Mask Language Modelling for generating potential word replacements in a black-box setting. Finally, Pruthi et al. [5] showed susceptibility of BERT [1] based models to character-level miss-spellings also in a black-box setting. In our study, we employ both character-level and word-level attacks in a white-box setting.

3 Problem Statement

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

4 Experimental Setting

4.1 Undefended Models

4.1.1 Datasets

We present our work based on HASOC 2019 (German Language) [6] and GermEval 2021 [7] sub-task 1 respectively. 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. Dataset statistics.

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

4.1.2 Training

For training, the undefended models, we fine-tune the 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 dataset respectively.

Table 2. Undefended models.

Dataset	Accuracy(%)
HASOC 2019	84
GermEval 2021	69

4.2 Attacks

4.2.1 Baseline Word-level White-Box Attack

The baseline word-level attack is composed by enhancing Hsieh et al. [3] 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 using *Masked Language Modeling* (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 as proposed by Garg et al. [4] we insert generated tokens to left/right of the target token where the candidate tokens are generated via MLM.

4.2.2 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 pre-trained 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. Firstly, we compute the document-level embeddings for both perturbed and source sequence 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. [4] 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 token should be same.

4.2.3 Character-level White-Box attack

In this white-box character level attack, similar to earlier white-box word-level attacks attention scores are obtained in order to get the word importance. Then, by ordering the word importance in the order of higher to lower we employ the character perturbation scheme employed by Pruthi et al. [5] since they evaluated this in the black-box setting only, we perform character level perturbation within a target token by token modification of character (swap, insert, delete etc) 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 the results section.

4.3 Defenses

4.3.1 Abstain Based Training

In several past evaluations and benchmarks of defenses against adversarial examples [10, 11, 12, 13, 14, 15], adversarial training [16] has been found to be one of the best ways 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, 18, 19, 20, 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 dataset (of C), where the adversarial examples have the abstain label, to create a new dataset. We then simply train on this dataset. We applied this defense strategy on the models from Sec. 4.1.2 and present the results in Table 4. We also show the classification attributions in Fig. 1 to try to interpret the models' behaviour.

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.)

4.3.2 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 character-level defense which is an unsupervised approach which makes an assumption that

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

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 worst case which means T_{input} is perturbed with character-level noise.

In this defense method, we firstly 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 datasets. However, we change input to character level by inputting word pairs to the network. Concretely, we labelled the Birkbeck spelling error corpus [23] which has word pairs with one correct and the other misspelled word and we label the each pair based on the Levenshtein distance between each pair. The schematics of our neural approach are given in Fig. 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.

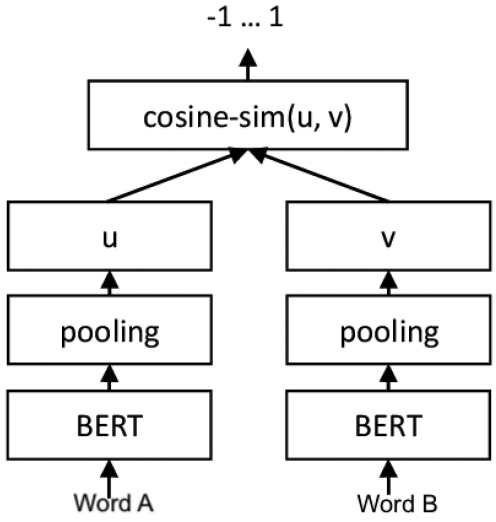

Figure 2. Sentence-BERT for character-level similarity.

Algorithm 1 Explicit Character Level Defense

$V_{train} \leftarrow t_1 \dots t_m$ ▷ Set of tokens in vocabulary
$E_v \leftarrow \vec{e_1} \dots \vec{e_m}$ ▷ Embeddings of vocabulary
$T_{input} \leftarrow t_1 \dots t_j$ ▷ Set of tokens in input
for $k \leftarrow 1$ to j **do**
  $\vec{e_k} \leftarrow v_1 \dots v_n$ ▷ Get embedding of input token k
  $\vec{scores} \leftarrow cos(E_v, \vec{e_k})$
    ▷ Cosine similarity with vocabulary embeddings
  if $\max \vec{scores} \geq 0.7$ *and* $\max \vec{scores} < 1.0$
then $vocab_{index} \leftarrow \arg\max \vec{scores}$;
    $T_{input}[k] \leftarrow V_{train}[vocab_{index}]$

end for

5 Results

5.1 Attack Results

As shown in Table 3 character-level attacks prove to be most effective on both models.

Table 3. Attacks result on undefended models.

Dataset	Attack	Success rate(%)
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 illustrates how number of queries required per sample for a successful attack depends on the dataset and

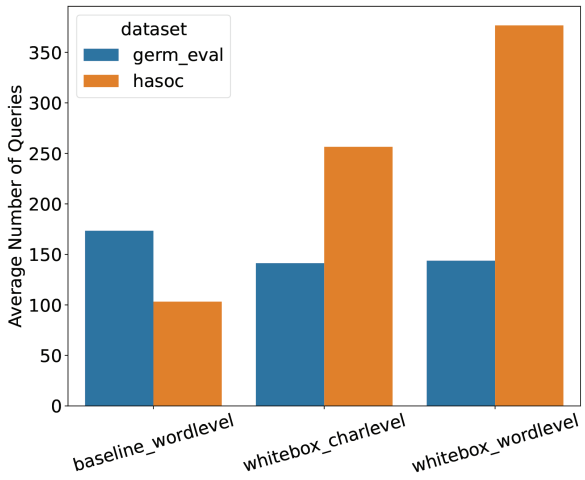

Figure 3. Average number of queries per successful attack.

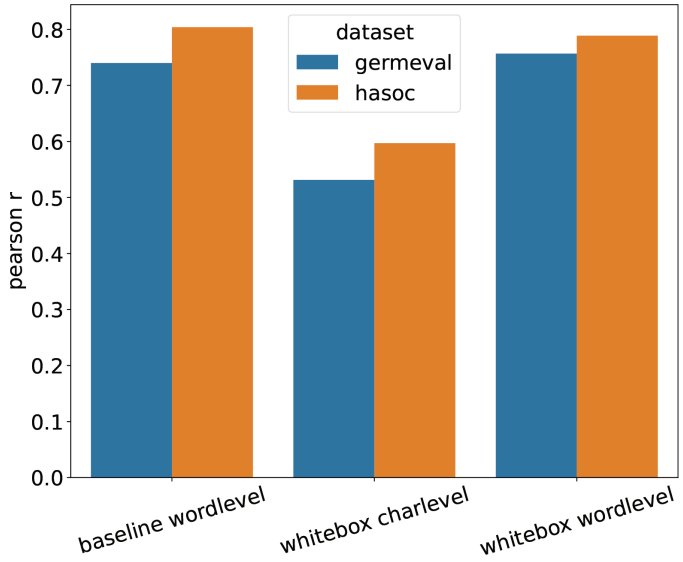

Figure 4. Pearson correlation between original text length and number of queries for attack success.

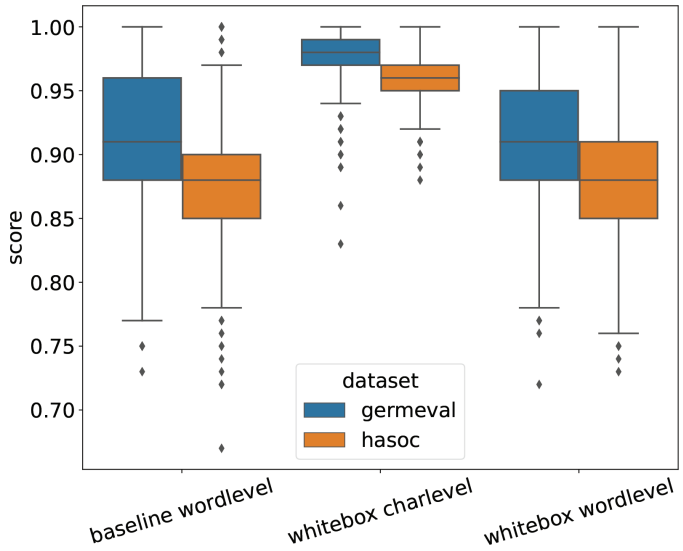

Figure 5. Levenshtein distance based similarity between original and perturbed sequences.

attack type, we further show in Fig. 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 require minimal amount of perturbation since the changes are at word level, moreover from Fig. 6 it can be concluded that character-level attack also makes the highest difference in model prediction confidence in case of a successful attack.

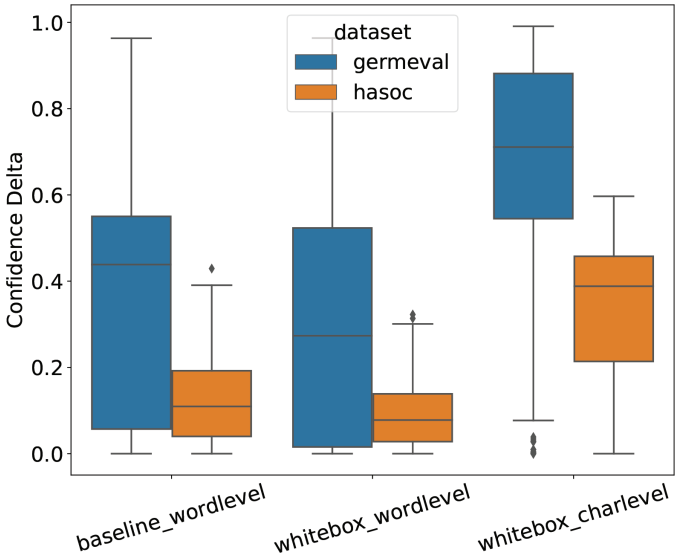

Figure 6. Confidence Delta between original and perturbed sequences caused by each attack.

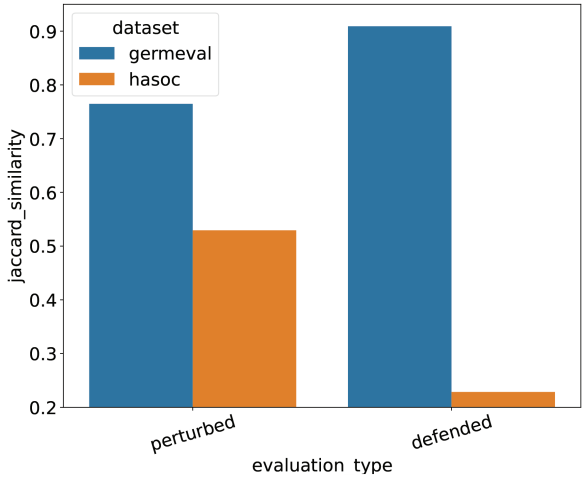

Figure 7. Jaccard similarity between original and perturbed text vs. the original and defended text.

5.2 Defense Results

Table 4. Character-level attack on defended models.

Dataset	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

6 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, since as shown in Fig. 7 current approach enhance 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 dataset 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 shown in Table 4.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[3] Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Siddhant Garg and Goutham Ramakrishnan. BAE: bert-based adversarial examples for text classification. *CoRR*, abs/2004.01970, 2020.

[5] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. *CoRR*, abs/1905.11268, 2019.

[6] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics, September 2021.

[8] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. *CoRR*, abs/2010.10906, 2020.

[9] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? natural language attack on text classification and entailment. *CoRR*, abs/1907.11932, 2019.

[10] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

[11] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[12] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *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*, pages 2206–2216. PMLR, 2020.

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *CoRR*, abs/2010.09670, 2020.

[15] Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini. Evading adversarial example detection defenses with orthogonal projected gradient descent. *CoRR*, abs/2106.15023, 2021.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

[17] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick D. McDaniel. On the (statistical) detection of adversarial examples. *CoRR*, abs/1702.06280, 2017.

[18] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *CoRR*, abs/1704.04960, 2017.

[19] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1563–1572. IEEE Computer Society, 2016.

[20] Angelo Sotgiu, Ambra Demontis, Marco Melis, Battista Biggio, Giorgio Fumera, Xiaoyi Feng, and Fabio Roli. Deep neural rejection against adversarial examples. *EURASIP Journal on Information Security*, 2020:1–10, 2020.

[21] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[23] Birkbeck spelling error corpus/roger mitton. Oxford Text Archive.