A Robust Method to Protect Text Classification Models against Adversarial Attacks

Bala Mallikarjunarao Garlapati, Ajeet Kumar Singh, Srinivasa Rao Chalamala
TCS Research, Tata Consultancy Services Ltd., India

Abstract
Text classification is one of the main tasks in natural language processing. Recently, adversarial attacks have shown a substantial negative impact on neural network based text classification models. There are few defenses to strengthen model predictions against adversarial attacks; popular among them are adversarial training and spelling correction. While adversarial training adds different synonyms to the training data, spelling correction methods defend against character variations at the word level. The diversity and sparseness of adversarial perturbations of different attack methods challenge these approaches. This paper proposes an approach to correct adversarial samples for text classification tasks. Our proposed approach combines grammar correction and spelling correction methods. In this, we use Gramformer for grammar correction and Textblob for spelling correction. These approaches are generic and can be applied to any text classification model without any retraining. We evaluated our approach with two state of the art attacks, DeepWordBug and TextBugger, on three open-source datasets IMDB, CoLA, and AG-News. The experimental results show that our approach can effectively counter adversarial attacks on text classification models while maintaining classification performance on original clean data.

Introduction
Text classification is the process of categorizing a text into one of a predefined set of topics. The text classification model tries to understand the sentence and classify them into appropriate categories. Toxic content classification in social media environments and automatic document classification are few of the many applications used in real world. Their usage in security-sensitive applications is increasing. For example, social media platforms (e.g., Facebook and Twitter) generate a vast amount of text every day, and some portions of it have harmful content, which is a threat to society. This toxic content propagation needs to be restricted by identifying and removing it.

DNN-based text classification models (Devlin et al. 2018) (Liu et al. 2019) outperform over traditional models, but they are sensitive to adversarial attacks because of their elemental mechanism. DNNs making wrong decisions by the addition of imperceptible perturbations to input samples is called an adversarial attack. Recent adversarial attack techniques (Gao et al. 2018) (Li et al. 2018) have shown a strong impact on fooling text classification model decisions. An adversary consciously creating attacks or similar mistakes made by humans (unintentionally) will effect same on the model. Figure 1 shows text classification model predictions against original text sample and adversarial sample. In this example, the adversary obfuscates some words with their modifications, replacing “fine” with “dramatic”, to make model decisions positive to negative. These attacks might cause security threats and loss of valuable information in commercial applications. Hence, Effective defense techniques are needed here to fight against such attacks.

Figure 1: Sentiment classification model predictions of original sample and adversarial sample

To defend against such attacks and improve the robustness of text classification models, present literature proposed techniques related to adversarial training (Wang et al. 2021), (Ebrahimi, Lowd, and Dou 2018) and spelling correction methods (Pruthi, Dhingra, and Lipton 2019), (Zhou et al. 2019). These methods have their own limitations. Our approach corrects the adversarial sentences and produces a supposedly clean output, which is required to obtain correct classification label and achieve good defense performance. Our contributions to this paper can be summarized as follows:

- We propose a defense approach which is a combination of spelling correction and grammar correction. This approach produces a high quality corrected output, useful to get original input classification labels for the corrected sentences of attacked samples.
- We evaluate our approach on three open-source text
classification datasets IMDB (Maas et al. 2011), AGNews (Zhang, Zhao, and LeCun 2015), and CoLA (Warstadt, Singh, and Bowman 2019). These datasets contain binary class and multi class data for text classification. And two different adversarial methods DeepWordBug and TextBugger used two generate adversarial samples on the above datasets.

- Our defense approach achieves accuracy of 92% on IMDB data, 80% on AGNews data and 76 % on CoLA data attacked samples. These results show that our approach has improved accuracy.

**Related Work**

A few existing attacks on text classification models in the literature are discussed here. (Miyato, Dai, and Goodfellow 2016) proposed work related to adversarial attacks on word embeddings. In this work, word embeddings are perturbed instead of input text sentences to fool a LSTM-based text classifier. (Liang et al. 2017) proposed a method to manipulate the input text at the word level. The adversary uses important phrases to detect failure scenarios and composes hot words to add, remove or insert in a meaningful manner. (Gao et al. 2018) proposed a black-box attack called DeepWordBug, which is based on the score of key tokens. In this method, four types of imperceivable perturbations are proposed to fool a text classifier. TextBugger (Li et al. 2018) is another black-box attack method, generated bugs using five different ways, and a new scoring method was introduced to improve the attack success rate. A genetic-based optimization method proposed by (Alzantot et al. 2018), which creates semantically similar adversarial samples to mislead the DNN-based text classification model. (Lei et al. 2018) proposed a gradient-based method to find valid paraphrases for given words using greedy search and provides an approximate factor for text classification models.

Few existing works available in the literature on defending adversarial attacks for text classification models. (Wang et al. 2021) proposed a method called SEM (Synonym Encoding Method) which used synonym word embeddings using an extra encoding stage in the training process to defend adversarial samples. It has a limitation of synonym substitution size. (Wang and Wang 2020) proposed the RSE (Random Substitution Encoder) method, which is a flexible synonym substitution encoding method to detect adversarial attacks. (Ebrahimi, Lowd, and Dou 2018) proposed a method to tackle adversarial samples by retraining the models with adversarial training data and showcased the improvement in model robustness. Another important defense category is spell checking for character-level attacked adversarial text. (Zhou et al. 2019) introduced a framework for adversarial text detection without retraining the model parameters. (Pruthi, Dhingra, and Lipton 2019) have designed a word recognition model to detect misspellings. They build a model upon the RNN semi-character architecture and are trained to identify corrupted words.

Our proposed approach overcomes the limitations of the SOTA methods, i.e., limited re-training data for adversarial training and limited character variations in spelling correction methods.

### Adversarial Sample Generation

Latest work on adversarial attacks against text classifiers has proved that imperceptibly achieved by modifying the input text’s spelling, words, or structure. To measure the strength of our proposed approach against adversarial attacks, we evaluated two SOTA adversarial attacks. Those are TextBugger (Li et al. 2018) and DeepWordBug (Gao et al. 2018), which have black-box access to the models and pose serious threats in real-world applications.

**DeepWordBug Attack:** DeepWordBug attack (Gao et al. 2018) modifies the characters in important words using various operations. In the first step, query the target model multiple times and calculate the word importance score using different scoring functions, Temporal Head Score (THS), Temporal Tail Score (TTS), and Combination Score to determine important tokens. In the second step, the important words with greater scores are modified at the character level (Insertion, Deletion, Swap, and Substitution) to generate adversarial sequences. **TextBugger Attack:** TextBugger (Li et al. 2018) can generate adversarial text under both black box and white box settings. In TextBugger attack first determines the classification confidence score of the important words. Then select a precisely crafted bug word to replace these important words to fool the text classification model.

Generated 1000 adversarial samples using above two attack methods, for three different text Classification Models. Corresponding results are shown in Table 1. DeepWordBug and TextBugger attack success rates are good on binary class datasets IMDB and CoLA. But for multi-class AGNews dataset, the attack success rate is medium. Successfully attacked samples are used for the evaluation of our approach. The average perturbed word rate is good for all the attacks, but not shown here because of space problem.

### Proposed Approach

We proposed a generic approach to correct adversarial samples in text classification applications. It is a combination of grammar correction and spelling correction modules and used Textblob (12) and Gramformer (PrithivirajDamodaran 2021) respectively for this purpose. The main advantage of our defense approach is that it does not involve any retraining procedure and does not require deep learning model’s structure modifications. TextBlob (12) is a textual data processing library and is used for spell corrections on input sentences. Textblob uses Peter Norvig’s work (spell correction 2007) for spelling correction. The main parts of this spelling correction strategy are the language model, error

| Attack Type       | IMDB Data | AGNews Data | CoLA Data |
|-------------------|-----------|-------------|-----------|
| DeepWordBug       | 79%       | 58%         | 91%       |
| TextBugger        | 92%       | 53%         | 85%       |

Table 1: Adversarial sample generation results on different datasets
model, and candidate model. The highest combined score of these models returns the most probable candidate word. Gramformer (Prithviraj Damodaran 2021) is a transformer-based deep generative model, that takes a distorted sentence as an input and correct grammatical errors in natural language text. It is mainly used to detect, highlight and correct grammar errors.

In this, first the perturbed sentences provided as input to Textblob, which recommends corrected words for misspelled words. The Textblob correction method considers the attack scenarios (edit distance 1 and 2) and finds the maximum probability of intended correction for the given original word. We added additional vocabulary in the text blob to improve defense accuracy. This additional vocabulary was provided by Juditacs, University of Washington (13) and the same is used for BERT models.

Gramformer is used to correct the grammar of the spell-corrected output sentences from the Textblob. The Quality estimator in this makes sure the corrections made are high quality and filters best probable candidate from top-N candidates. We overcome the limitation of the Gramformer model’s support for max length by dividing the input text into a maximum of 128 length word sentences. After processing it through Gramformer, we combined all corrected sentences for model prediction. Our approach produces high quality sentences for adversarial samples, as Textblob produces maximum probable correction word for the given adversarial word (out of all possible candidate words) and Gramformer produces corrected sentences with help of quality score to filter from best top-N candidates. In this way, our approach helps in obtaining the corrected sentence label as the original input label for a given text classification model. Experimental results in the next section show the effectiveness of our proposed approach on attacked samples and clean samples as well.

**Experimental Results**

In this section, we examine the performance of the proposed approach by conducting extensive experiments. We evaluated the proposed approach on three popular data sets IMDB (Maas et al. 2011), AGNews (Zhang, Zhao, and LeCun 2015) and CoLA (Warstadt, Singh, and Bowman 2019). First, adversarial text is generated for the above three datasets using two different attacks methods DeepWordBug and TextBugger. Only the samples which are successful in fooling the text classification are used for evaluation of the propose approach. We evaluated the robustness of three different text classification models on Gramformer. Next, we evaluated defense accuracy on our proposed approach, which is a combination of Textblob and Gramformer on adversarial samples. For this we used textattack framework for applying adversarial attacks and other NLP tasks.

Table 2 shows empirical results of Gramformer defense and the proposed approach against DeepWordBug and TextBugger attacks on IMDB data. The BERT based uncased model (Devlin et al. 2018) is used for text classification and is fine-tuned for the IMDB dataset. ‘Gramformer only’ method gave 90% accuracy on attacked IMDB samples and 98% accuracy on original clean samples. And our proposed approach (Textblob+Gramformer) gave 92% accuracy on attacked IMDB samples and nearly 97% accuracy on original clean samples.

Table 3 shows the Gramformer defense and the proposed approach defense results against DeepWordBug and TextBugger attacks on AgNews data. The Roberta base model (Liu et al. 2019) was fine-tuned for text classification on AgNews dataset. ‘Gramformer only’ method gave 77% accuracy on attacked AgNews samples and 90% accuracy on original clean samples. Our proposed approach (Textblob+Gramformer) gave 77% accuracy on attacked AgNews samples and nearly 92% accuracy on original clean samples.

Table 4 shows the Gramformer defense and the proposed approach defense results against DeepWordBug and TextBugger attacks on CoLA data. The Bert base uncased model (Devlin et al. 2018) is used for text classification and is fine-tuned for the CoLA dataset. ‘Gramformer only’ method gave more than 73% accuracy on attacked CoLA samples and 91% accuracy on original clean samples. Our approach (Textblob+Gramformer) gave 76% accuracy on attacked CoLA samples and nearly 90% accuracy on original clean samples. We can observe that there is an increase in the proposed defense accuracy on the three dataset attacked samples compared to Gramformer only.

(Zhou et al. 2019) proposed Discriminate Perturba-
| Attack Type         | Gramformer | TextBlob + Gramformer |
|---------------------|------------|-----------------------|
| DeepWordBug         | 74.39%     | 81.00%                |
| TextBugger          | 72.63%     | 71.32%                |
| No Attack (Clean Samples) | 91.10%     | 89.49%                |

Table 4: Proposed Approach Results On CoLA Dataset

Conclusions and Future work

We proposed a defense mechanism for adversarial attacks against text classification tasks. The proposed approach is a generic method, and can be used to correct adversarial text for any type of text classification models. This proposed strategy which is a combination of Gramformer and Textblob (spelling correction and grammar correction) produce high-quality corrected text, against adversarial samples to obtain the correct labels. Our experimental results show that our approach is more effective to defend against various adversarial attacks. This can be attributed to the joint correction mechanism both at word and sentence levels. We plan to extend our work for other NLP tasks such as neural machine translation (NMT).

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating natural language adversarial examples. arXiv preprint arXiv:1804.07998.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ebrahimi, J.; Lowd, D.; and Dou, D. 2018. On adversarial examples for character-level neural machine translation. arXiv preprint arXiv:1806.09030.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), 50–56. IEEE.

Lei, Q.; Wu, L.; Chen, P.-Y.; Dimakis, A. G.; Dhillon, I. S.; and Witbrock, M. 2018. Discrete attacks and submodular optimization with applications to text classification. arXiv preprint arXiv:1812.00151.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2018. Textbugger: Generating adversarial text against real-world applications. arXiv preprint arXiv:1812.05271.

Liang, B.; Li, H.; Su, M.; Bian, P.; Li, X.; and Shi, W. 2017. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 142–150.

Miyato, T.; Dai, A. M.; and Goodfellow, I. 2016. Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.

PrithvirajDamodaran. 2021. Gramformer. https://github.com/PrithvirajDamodaran/Gramformer/. [Online; ].

Pruthi, D.; Dhingra, B.; and Lipton, Z. C. 2019. Combating adversarial misspellings with robust word recognition. arXiv preprint arXiv:1905.11268.

spell correction, N. 2007. Norvig spell correction. https://norvig.com/spell-correct.html. [Online; ].

Wang, Z., and Wang, H. 2020. Defense of word-level adversarial attacks via random substitution encoding. In International Conference on Knowledge Science, Engineering and Management, 312–324. Springer.

Wang, X.; Hao, J.; Yang, Y.; and He, K. 2021. Natural language adversarial defense through synonym encoding. In Uncertainty in Artificial Intelligence, 823–833. PMLR.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2019. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics 7:625–641.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. Advances in neural information processing systems 28.

Zhou, Y.; Jiang, J.-Y.; Chang, K.-W.; and Wang, W. 2019. Learning to discriminate perturbations for blocking adversarial attacks in text classification. arXiv preprint arXiv:1909.03084.