Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples

Maximilian Mozes[1] **Pontus Stenetorp**[1] **Bennett Kleinberg**[2,1] **Lewis D. Griffin**[1]
[1]University College London
[2]Tilburg University
{m.mozes, p.stenetorp, l.griffin}@cs.ucl.ac.uk
bennett.kleinberg@tilburguniversity.edu

Abstract

Recent efforts have shown that neural text processing models are vulnerable to adversarial examples, but the nature of these examples is poorly understood. In this work, we show that adversarial attacks against CNN, LSTM and Transformer-based classification models perform word substitutions that are identifiable through frequency differences between replaced words and their corresponding substitutions. Based on these findings, we propose *frequency-guided word substitutions* (FGWS), a simple algorithm exploiting the frequency properties of adversarial word substitutions for the detection of adversarial examples. FGWS achieves strong performance by accurately detecting adversarial examples on the SST-2 and IMDb sentiment datasets, with F_1 detection scores of up to 91.4% against RoBERTa-based classification models. We compare our approach against a recently proposed perturbation discrimination framework and show that we outperform it by up to 13.0% F_1.

1 Introduction

Artificial neural networks are vulnerable to adversarial examples—carefully crafted perturbations of input data that lead a learning model into making false predictions (Szegedy et al., 2014).

While initially discovered for computer vision tasks, natural language processing (NLP) models have also been shown to be oversensitive to adversarial input perturbations for a variety of tasks (Papernot et al., 2016; Jia and Liang, 2017; Belinkov and Bisk, 2018; Glockner et al., 2018; Iyyer et al., 2018). Here we focus on highly successful synonym substitution attacks (Alzantot et al., 2018; Ren et al., 2019; Zang et al., 2020), in which individual words are replaced with semantically similar ones. Existing defense methods against these attacks mainly focus on adversarial training (Jia and Liang, 2017; Ebrahimi et al., 2018; Ribeiro et al., 2018; Ren et al., 2019; Jin et al., 2019) and hence typically require a priori attack knowledge and models to be retrained from scratch to increase their robustness. Recent work by Zhou et al. (2019) instead proposes DISP (*learning to discriminate perturbations*), a perturbation discrimination framework that exploits pre-trained contextualized word representations to detect and correct word-level adversarial substitutions without having to retrain the attacked model. In this paper, we show that we can achieve an improved performance for the detection and correction of adversarial examples based on the finding that various word-level adversarial attacks have a tendency to replace input words with less frequent ones.[1] Figure 1 illustrates this tendency for two state-of-the-art attacks. We provide statistical evidence to support this observation and propose a rule-based and model-agnostic algorithm, *frequency-guided word substitutions* (FGWS), to detect adversarial sequences

Attack	Original or perturbed sequence
None	A clever blend of fact and fiction
GENETIC	A **brainy** [***clever***] blend of fact and fiction (1.39 ←-- 5.55)
PWWS	A **cunning** [***clever***] **blending** [***blend***] of fact and **fabrication** [***fiction***] (1.61 ←-- 5.55; 0.00 ←-- 3.81; 0.00 ←-- 4.39)

Figure 1: Corpus $\log_e$ frequencies of the replaced words (bold, italic, red) and their corresponding adversarial substitutions (bold, black) using the GENETIC (Alzantot et al., 2018) and PWWS (Ren et al., 2019) attacks on SST-2 (Socher et al., 2013).

[1]This frequency difference is expected for attacks that explicitly conduct symbol substitutions resulting in out-of-vocabulary (OOV) terms (Gao et al., 2018). We therefore study attacks that do not explicitly enforce a mapping to words that have lower frequencies.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 171–186
April 19 - 23, 2021.

and recover model performances for perturbed test set sequences. FGWS effectively detects adversarial perturbations, achieving F_1 scores of up to 91.4% against RoBERTa-based models (Liu et al., 2019) on the IMDb sentiment dataset (Maas et al., 2011). Furthermore, our results show that FGWS outperforms DISP by up to 13.0% F_1 when differentiating between unperturbed and perturbed sequences, despite representing a conceptually simpler approach to this task.

2 Generating adversarial examples

In our experiments, we investigate two baseline attacks introduced by Ren et al. (2019) as well as two state-of-the-art attacks.

RANDOM. Our first baseline attack is a simple word substitution model that randomly selects words in an input sequence and replaces them with synonyms randomly sampled from a set of synonyms related to the specific word. We follow Ren et al. (2019) by using WORDNET (Fellbaum, 1998) to identify synonym substitutions for each selected word.

PRIORITIZED. Our second baseline builds upon RANDOM by selecting the replacement word from the synonym set that maximizes the change in prediction confidence for the true label of an input.

GENETIC. We additionally analyze an attack suggested by Alzantot et al. (2018), consisting of a population-based black-box mechanism based on genetic search that iteratively performs individual word-level perturbations to an input sequence to cause a misclassification.

PWWS. Lastly, we analyze the *probability weighted word saliency* (PWWS) algorithm (Ren et al., 2019). For each word in an input sequence, PWWS selects a set of synonym replacements from WORDNET and chooses the synonym yielding the highest difference in prediction confidence for the true class label after replacement. The algorithm furthermore computes the word saliency (Li et al., 2016a,b) for each input word and ranks word replacements based on these two indicators.

Datasets and models. We perform experiments on two binary sentiment classification datasets, the *Stanford Sentiment Treebank* (SST-2, Socher et al., 2013) and the IMDb reviews dataset (Maas et al., 2011), both of which are widely used in related work focusing on adversarial examples in NLP (Jia et al., 2019; Ren et al., 2019; Zhou et al., 2019). Dataset details can be found in Appendix A. Adhering to Zhou et al. (2019), we attack a pretrained model based on the Transformer architecture (Vaswani et al., 2017). Zhou et al. (2019) use BERT (Devlin et al., 2019) in their experiments, but we found that RoBERTa (Liu et al., 2019) represents a stronger model for the specified tasks.

We additionally experiment with both a CNN (Kim, 2014) and an LSTM (Hochreiter and Schmidhuber, 1997) text classification model, both of which have been employed in existing work studying textual adversarial attacks (Alzantot et al., 2018; Lei et al., 2019; Jia et al., 2019; Tsai et al., 2019; Ren et al., 2019).

The fine-tuned RoBERTa model achieves 93.4% and 94.9% accuracy on the IMDb and SST-2 test sets, which is comparable to existing work (Beltagy et al., 2020; Liu et al., 2019). On the IMDb test set, the CNN achieves an accuracy of 86.0% and the LSTM achieves 83.1%. These performances are close to existing work using comparable settings (Zhang et al., 2019; Ren et al., 2019). On the SST-2 test set, the CNN achieves 84.0% and the LSTM 85.2% accuracy, which are also close to comparable experiments (Huang et al., 2019).

Following Ren et al. (2019), we apply all four attacks to a random subset of 2,000 sequences from the IMDb test set as well as the entire test set of SST-2 (1,821 samples). Implementation details for the models and attacks can be found in Appendix B. We report the after-attack accuracies[2] for the RoBERTa model in Table 2 and for the CNN/LSTM models in Table 3 (column Adv.). We observe that all four attacks cause notable decreases in model accuracy on the test sets, and that GENETIC and PWWS are more successful than the baseline attacks in most comparisons.

3 Analyzing frequencies of adversarial word substitutions

Next, we conduct an analysis of the word frequencies of individual words replaced by the attacks and their substitutions. We compute the $\log_e$ training set frequencies $\phi(x)$ of all words x that have been replaced by the respective attacks and all of their corresponding substitutions. Then, we conduct Bayesian hypothesis testing (Rouder et al., 2009) to statistically compare the two samples. This is achieved by computing the Bayes factor BF_{10}, rep-

[2]The after-attack accuracy represents the model accuracy on the test set after perturbing all correctly classified inputs. A lower after-attack accuracy indicates a stronger attack.

Dataset	Attack	Replaced		Subst.			non-OOV		
		μ_ϕ	σ_ϕ	μ_ϕ	σ_ϕ	d	μ_ϕ	σ_ϕ	d
IMDb	RANDOM	7.6	2.5	3.4	2.8	**1.6**	4.4	2.4	1.3
	PRIORITIZED	7.6	2.5	3.6	2.8	1.5	4.4	2.4	1.3
	GENETIC	6.5	2.0	3.7	2.3	1.3	4.0	2.2	1.2
	PWWS	6.9	2.3	4.4	2.5	1.0	5.0	2.1	0.9
SST-2	RANDOM	5.4	2.6	2.1	2.4	**1.4**	4.0	1.8	0.6
	PRIORITIZED	5.4	2.6	2.1	2.4	1.3	4.0	1.8	0.6
	GENETIC	4.4	1.9	1.9	2.2	1.2	3.6	1.6	0.4
	PWWS	4.8	2.1	2.9	2.2	0.9	4.0	1.5	0.4

Table 1: Mean $\log_e$ frequencies of replaced words and their substitutions. Values in bold denote largest effect sizes per dataset.

resenting the degree to which the data favor the alternative hypothesis over the null hypothesis. Here, the alternative hypothesis $\mathcal{H}_1$ states that *the frequencies of replaced words differ from the frequencies of the adversarial substitutions*. The null hypothesis $\mathcal{H}_0$ states that *there is no such difference*. The higher BF_{10}, the stronger the evidence in favor of the alternative hypothesis $\mathcal{H}_1$.[3] We additionally calculate Cohen's d effect sizes for all mean frequency comparisons.[4]

Table 1 shows the $\log_e$ frequencies (mean μ_ϕ and standard deviation σ_ϕ) and Cohen's d for the specified samples generated by the attacks against the RoBERTa model (the results for the CNN and LSTM models can be found in Appendix C). We report the mean frequencies of all adversarial substitutions (Subst.) and only those that occur in the training set (non-OOV), to demonstrate that the frequency differences are not solely caused by OOV substitutions. Across datasets and attacks, the substitutions are consistently less frequent than the words selected for replacement. We observe large Cohen's d effect sizes for the majority of comparisons, statistically supporting the observation of mean frequency differences between replaced words and their corresponding substitutions. We furthermore observe that $\mathrm{BF}_{10} > 10^{55}$ holds for all comparisons—both when considering all and only non-OOV substitutions (the BF_{10} scores can be found in Appendix D). This provides strong empirical evidence that $\mathcal{H}_1$ is more likely to be supported by the measured word frequencies (see Appendix E for additional illustrations).

[3] A Bayes factor $\mathrm{BF}_{10} > 100$ can be interpreted as "extreme" evidence for $\mathcal{H}_1$ (Wagenmakers et al., 2011).

[4] Cohen's d indicates the magnitude of the frequency differences of the two samples—larger effect sizes suggest a higher magnitude of the frequency difference. A value of $d = 0.8$ can be interpreted as a large effect, $d = 0.5$ is considered a moderate effect (Cohen, 1988).

4 Frequency-guided word substitutions

Based on the observation of consistent frequency differences between replaced words and adversarial substitutions, we argue that the effects of such substitutions can be mitigated through simple frequency-based transformations. To do this, we propose *frequency-guided word substitutions* (FGWS), a detection method that estimates whether a given input sequence is an adversarial example.[5] We denote a classification model by a function $f(X)$ that maps a sequence X to a C-dimensional vector representing the probabilities for predicting each of the C possible classes. We represent a sequence as $X = \{x_1, \ldots, x_n\}$, where x_i denotes the i-th word in the sequence. We furthermore introduce the notation $f^*(X) \in \{1, \ldots, C\}$ representing the class label predicted by f given input X. FGWS transforms a given sequence X into a sequence X' by replacing infrequent words with more frequent, semantically similar substitutions. We initially define the subset $X_E := \{x \in X \mid \phi(x) < \delta\}$ of words that are eligible for substitution, where $\delta \in \mathbb{R}_{>0}$ is a frequency threshold. FGWS then generates a sequence X' from X by replacing all eligible words with words that are semantically similar, but have higher occurrence frequencies in the model's training corpus. For each eligible word $x \in X_E$ we consider the set of replacement candidates $\mathcal{S}(x)$ and find a replacement x' by selecting $x' = \mathrm{argmax}_{w \in \mathcal{S}(x)}\ \phi(w)$. We then generate X' by replacing each eligible word x with x' if $\phi(x') > \phi(x)$. Given the prediction label $y = f^*(X)$ for X and a threshold $\gamma \in [0, 1]$, the sequence X is considered adversarial if $f(X)_y - f(X')_y > \gamma$, i.e., if the difference in prediction confidence on class y before and after transformation exceeds the threshold γ. The threshold allows control of the rate of false positives (i.e., unperturbed sequences that are erroneously identified as adversarial) flagged by our method.

4.1 Comparisons

DISP. We compare FGWS to the DISP framework (Zhou et al., 2019), which is, to the best of our knowledge, the best existing approach for the detection of word-level adversarial examples. DISP uses two independent BERT-based components, a perturbation discriminator and an embedding estimator for token recovery, to identify perturbed

[5] Code is available at https://github.com/maximilianmozes/fgws.

Dataset	Attack	Adv.	Restored acc.		TPR (FPR)		F_1	
			DISP	FGWS	DISP	FGWS	DISP	FGWS
IMDb	RANDOM	87.3	89.2	**91.0**	63.6 (9.4)	**83.5** (9.3)	73.6	**86.6**
	PRIORITIZED	41.5	81.0	**85.9**	87.8 (9.4)	**92.0** (9.3)	89.0	**91.4**
	GENETIC	47.7	74.1	**80.6**	70.4 (9.4)	**81.5** (9.3)	78.3	**85.4**
	PWWS	41.0	68.7	**75.4**	66.2 (9.4)	**76.4** (9.3)	75.4	**82.3**
SST-2	RANDOM	87.2	86.6	**90.0**	**66.2** (11.9)	61.3 (11.4)	**74.4**	71.0
	PRIORITIZED	68.9	80.8	**84.8**	69.1 (11.9)	**74.7** (11.4)	76.3	**80.3**
	GENETIC	40.8	60.1	**61.7**	**57.2** (11.9)	57.0 (11.4)	**67.7**	**67.7**
	PWWS	57.4	71.0	**78.2**	59.6 (11.9)	**65.6** (11.4)	69.6	**74.2**

Table 2: Adversarial example detection performances for DISP and FGWS when evaluated on attacks against RoBERTa. Adv. shows the model's classification accuracy on the perturbed sequences. Restored acc. denotes model accuracy on the adversarial sequences after transformation. Values in bold represent best scores per metric, dataset and attack.

Model/ Dataset	Attack	Adv.	Restored acc.		F_1	
			NWS	FGWS	NWS	FGWS
CNN/ IMDb	RANDOM	73.0	79.5	**84.7**	75.2	**83.5**
	PRIORITIZED	14.0	41.6	**78.9**	71.5	**89.3**
	GENETIC	10.7	21.3	**68.5**	37.9	**83.5**
	PWWS	10.2	27.4	**70.2**	45.4	**83.9**
LSTM/ IMDb	RANDOM	64.7	75.7	**80.9**	80.5	**83.9**
	PRIORITIZED	3.2	32.0	**71.6**	62.4	**86.6**
	GENETIC	1.2	10.9	**54.9**	34.3	**78.0**
	PWWS	1.6	17.3	**57.1**	41.7	**77.4**
CNN/ SST-2	RANDOM	71.8	77.1	**78.4**	**71.4**	69.2
	PRIORITIZED	50.3	60.1	**69.3**	54.8	**67.8**
	GENETIC	19.6	34.9	**48.8**	49.9	**60.3**
	PWWS	28.1	47.4	**58.1**	55.1	**63.9**
LSTM/ SST-2	RANDOM	73.4	79.3	**80.5**	**69.2**	62.2
	PRIORITIZED	48.5	59.9	**74.0**	54.9	**67.3**
	GENETIC	21.3	37.6	**61.1**	51.2	**62.8**
	PWWS	28.6	49.7	**67.2**	55.9	**63.4**

Table 3: Performance results of NWS and FGWS on attacks against the CNN and LSTM models. Values in bold indicate best performances per model-dataset-attack combination and metric.

tokens and to reconstruct the replaced ones.

NWS. For the CNN and LSTM models, we compare FGWS with the *naive word substitutions* (NWS) baseline. For a given input sequence, NWS selects all OOV words in that sequence and replaces each with a random choice from a set of semantically related words. We restrict NWS to allow only substitutions for which the replacement word occurs in the model's training vocabulary. NWS can be interpreted as a variant of FGWS that is not explicitly guided by word frequencies.

4.2 Experiments

We apply both methods to the adversarial examples crafted by the four attacks on the subsets of both the IMDb and SST-2 datasets as described in Section 2. To account for an imbalance between unperturbed and perturbed sequences, we repeatedly bootstrap a balanced set of unperturbed sequences for each set of perturbed sequences for 10,000 times and compute the average detection scores. For FGWS, we tune the frequency threshold δ for each model-dataset combination on the validation set. To do this, we utilize the PRIORITIZED attack to craft adversarial examples from all sequences of the validation set[6] and compare FGWS detection performances with different values for δ. Specifically, we set δ equal to the $\log_e$ frequency representing the q^{th} percentile of all $\log_e$ frequencies observed by the words eligible for replacement in the training set, and experiment with $q \in \{0, 10, \ldots, 100\}$. We select γ so that not more than 10% of the unperturbed sequences in the validation set are labeled as adversarial.[7] For FGWS, we define the set of replacement candidates for each word $x \in X_E$ as the union of the word's K nearest neighbors in a pre-trained GLOVE (Pennington et al., 2014) word embedding space and its synonyms in WORDNET. We set K equal to the average number of WORDNET synonyms for each word in the validation set (yielding $K = 6$ for IMDb and $K = 8$ for SST-2).

4.3 Results

We report the results comparing FGWS to DISP on attacks against RoBERTa in Table 2. Here, the true positive rate (TPR) represents the percentage of successful adversarial examples that were cor-

[6] We assume both baseline attacks as given to the defender, and prefer PRIORITIZED over RANDOM due to increased effectiveness and hence a larger sample size for parameter tuning.

[7] We provide additional results with varying false positive thresholds in Appendix F.

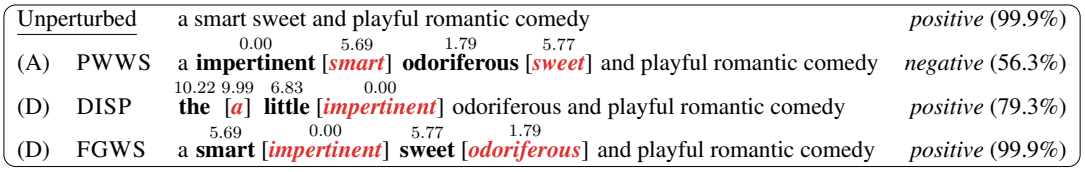


Figure 2: The detection methods applied to an adversarial example from the PWWS attack against RoBERTa on SST-2. The words highlighted in bold, italic and red were selected for replacement by the attack (A) and the detection methods (D), the ones in bold and black denote the substitutions. The values above the words denote their $\log_e$ frequencies.

rectly identified as such, and the false positive rate (FPR) denotes the percentage of unperturbed sequences that were identified as adversarial. The column Adv. gives the classification accuracy on the perturbed sequences, and Restored acc. the model's accuracy on the adversarial sequences after transformation. We observe that FGWS best restores the model's classification accuracy across all comparisons, showing it to be effective in mitigating the effects of the individual attacks. Furthermore, FGWS outperforms DISP in terms of true positive rates and F_1 across the majority of experiments. These results show that, although contextualized word representations (DISP) serve as a competitive method to detect adversarial examples, relying solely on frequency-guided substitutions (FGWS) shows to be more effective. Figure 2 provides an example adversarial sequence generated with the PWWS attack and the two corresponding transformed sequences using DISP and FGWS (see Appendix G for additional examples).

The results of NWS and FGWS against the CNN and LSTM models are shown in Table 3. We observe that FGWS outperforms NWS across all comparisons in terms of restored model accuracy and in the majority of comparisons in terms of F_1. Moreover, the direct comparison between NWS and FGWS again underlines the importance of utilizing word frequencies as guidance for the word substitutions: while NWS is not guided by word frequency characteristics to perform the word replacements, we observe that FGWS outperforms NWS by a large margin in most comparisons, demonstrating the effectiveness of mapping infrequent words to their most frequent semantically similar counterparts to detect adversarial examples.

4.4 FGWS on unperturbed data

We furthermore investigate the effect of FGWS on model performance on unperturbed sequences after transformation. To do this, we transform the sampled test sets using FGWS and evaluate classification accuracies after sequence transformation. The differences in accuracy for the CNN, LSTM and RoBERTa models before and after transformation are 0.0%, $+1.0\%$ and -0.2% for IMDb and -1.8%, -2.9% and -1.8% for SST-2. This indicates that FGWS applied to unperturbed data has only small effects on classification accuracy, and in some cases even slightly increases prediction accuracy.

5 Limitations

It is worth mentioning that compared to FGWS, DISP represents a more general perturbation discrimination approach since it is trained to detect both character- and word-level adversarial perturbations, whereas FGWS solely focuses on word-level attacks.

Furthermore, it remains open whether FGWS would be effective against attacks for which the frequency difference is less evident. To investigate this, we conducted preliminary experiments by restricting the investigated attacks to only allow equifrequent substitutions. However, we observed that introducing this constraint has a substantial effect on attack performance, since the attacks are supplied with fewer candidate replacements. We will further investigate this in future work.

6 Conclusion

We have shown that the word frequency characteristics of adversarial word substitutions can be leveraged effectively to detect adversarial sequences for neural text classification. Our proposed approach outperforms existing detection methods despite representing a conceptually simpler approach to this task.

Acknowledgements

This research was supported by the Dawes Centre for Future Crime at University College London.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. Technical report, Google.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Academic press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

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

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4081–4091, Hong Kong, China. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4120–4133, Hong Kong, China. 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. *arXiv preprint arXiv:1907.11932*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G Dimakis, Inderjit S Dhillon, and Michael Witbrock. 2019. Discrete adversarial attacks and submodular optimization with applications to text classification. In *SysML 2019*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Nicolas Papernot, Patrick Drew McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016 - 2016 IEEE Military Communications Conference*, Proceedings - IEEE Military Communications Conference MILCOM, pages 49–54, United States. Institute of Electrical and Electronics Engineers Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, 16(2):225–237.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Yi-Ting Tsai, Min-Chu Yang, and Han-Yu Chen. 2019. Adversarial attack on sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 233–240, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han LJ Van Der Maas. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011). *Journal of Personality and Social Psychology*, 100(3):426 – 432.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Yichao Zhou, Jyun-Yu Jiang, Kai-Wei Chang, and Wei Wang. 2019. Learning to discriminate perturbations for blocking adversarial attacks in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4903–4912, Hong Kong, China. Association for Computational Linguistics.

A Dataset statistics

The SST-2 dataset comes with a pre-defined split of 67,349 samples for training, 872 for validation and 1,821 for testing. The IMDb dataset consists of 50,000 positive and negative movie reviews with a pre-defined split of 25,000 training and 25,000 test samples. Since this dataset does not have a pre-defined validation set, we hold out 1,000 randomly selected training set samples for validation. We select a validation set of roughly the same size as for SST-2 for fair comparisons when tuning parameters for adversarial example detection. To the best of our knowledge, the compared work (Alzantot et al., 2018; Ren et al., 2019) does not validate model performance on held-out training data.

B Model and attack details

B.1 RoBERTa

We utilize a pre-trained RoBERTa (base) model (Liu et al., 2019) provided by the *Hugging Face Transformers* library (Wolf et al., 2019). We use maximum input sequence lengths of 256 and 128 after byte-pair encoding (Sennrich et al., 2016) for the IMDb and SST-2 datasets, respectively. The RoBERTa model consists of 125 million parameters.[8] The model was trained for 10 epochs with batch size 32 (SST-2) and 16 (IMDb) and a learning rate of $1 \cdot 10^{-5}$. We evaluated model performance after each epoch on the validation set and selected the best-performing checkpoints for testing.

[8] https://github.com/pytorch/fairseq/tree/master/examples/roberta

B.2 CNN/LSTM

The CNN architecture consists of 3 convolutional layers with kernel sizes 2, 3 and 4 and 100 feature maps for each convolutional layer. The LSTM operates on a hidden state size of 128. Following Alzantot et al. (2018), we initialize the LSTM with pre-trained GLOVE (Pennington et al., 2014) word embeddings, and do the same for the CNN.

Both the LSTM and the CNN use *Dropout* (Srivastava et al., 2014) during training with a rate of 0.1 before applying the output layer. We trained both models for 20 epochs using the *Adam* optimizer (Kingma and Ba, 2014). We evaluated model performance after each epoch on the validation set and selected the best-performing checkpoints for testing. The CNN and LSTM models were trained with batch size 100 and a learning rate of $1 \cdot 10^{-3}$.

B.3 PWWS

Our implementation of PWWS is based on the code as provided by Ren et al. (2019) on GitHub.[9]

B.4 GENETIC

Note that we utilize a different language model for the `Perturb` subroutine as compared to the original implementation by Alzantot et al. (2018). While Alzantot et al. (2018) employ the Google 1 billion words language model (Chelba et al., 2013), we instead utilize the recently proposed GPT-2 language model (Radford et al., 2019) and compute the sequences' perplexity scores using the exponentialized language modelling loss (we employ the pre-trained `GPT2LMHeadModel` language model from Wolf et al. (2019)). We compute the perplexity scores for each perturbed sequence only around the respective replacement words by only considering a subsequence ranging from five words before to five words after an inserted replacement. The motivation for using a different language model as compared to the original implementation is due to computational efficiency, since we observed a notable decrease in attack runtime with our modification. This does not have an impact on attack performance, since our implementation of the GENETIC has an attack success rate of 98.6% against the LSTM on IMDb, whereas Alzantot et al. (2018) report an attack success rate of 97%.

For attacks against SST-2, we furthermore increase the δ threshold for the maximum distance between replaced words and substitutions to $\delta = 1.0$,

[9] https://github.com/JHL-HUST/PWWS

Dataset	Model	Attack	Replaced		Subst.			non-OOV		
			μ_ϕ	σ_ϕ	μ_ϕ	σ_ϕ	d	μ_ϕ	σ_ϕ	d
IMDb	CNN	RANDOM	7.6	2.5	3.5	2.8	**1.6**	4.4	2.4	1.3
		PRIORITIZED	7.6	2.5	3.3	2.7	**1.6**	3.9	2.5	1.5
		GENETIC	6.3	2.0	3.5	2.2	1.3	3.7	2.1	1.3
		PWWS	6.7	2.3	4.0	2.4	1.1	4.5	2.1	1.0
	LSTM	RANDOM	7.6	2.5	3.5	2.8	1.6	4.4	2.4	1.3
		PRIORITIZED	7.6	2.5	2.8	2.3	**2.0**	3.2	2.2	1.8
		GENETIC	6.2	2.0	3.1	1.9	1.6	3.3	1.8	1.5
		PWWS	6.4	2.2	3.5	2.1	1.4	3.7	1.9	1.3
SST-2	CNN	RANDOM	5.4	2.5	2.0	2.3	**1.4**	3.8	1.7	0.7
		PRIORITIZED	5.4	2.5	2.4	2.1	1.3	3.5	1.7	0.8
		GENETIC	4.3	1.8	2.2	1.9	1.2	3.2	1.4	0.6
		PWWS	4.8	2.1	2.8	2.1	1.0	3.8	1.5	0.6
	LSTM	RANDOM	5.4	2.6	2.0	2.3	**1.4**	3.8	1.7	0.7
		PRIORITIZED	5.4	2.5	2.3	2.1	1.3	3.4	1.6	0.9
		GENETIC	4.3	1.7	2.0	1.9	1.3	3.1	1.4	0.8
		PWWS	4.8	2.0	2.7	2.1	1.0	3.7	1.4	0.6

Table 4: Mean $\log_e$ frequencies of replaced words and their corresponding substitutions by attack, model and dataset. The shown values are the mean μ_ϕ and standard deviation σ_ϕ of the $\log_e$ frequencies corresponding to each setting, and additionally the Cohen's d effect sizes for the substitutions. Values in bold denote largest effect sizes per dataset and model.

since we observed poor attack performances with $\delta = 0.5$ (which was used by Alzantot et al. (2018) and in our experiments on IMDb). All other parameters of the attack (e.g., the number of generations and population size) are directly adapted from Alzantot et al. (2018).

We restrict the words eligible for replacement by the GENETIC attack to non-stopwords, in accordance to Alzantot et al. (2018). Since the attack computes nearest neighbors for a selected word from a pre-trained embedding space, we furthermore can only select words for which there exists an embedding representation in this pre-trained space. On the SST-2 test set, we found three input sequences consisting of only one word which we excluded from our evaluation, since the used GPT-2 language model implementation requires an input sequence consisting of more than one word.

B.5 RANDOM, PRIORITIZED, PWWS, GENETIC

For the GENETIC attack, we follow Alzantot et al. (2018) by limiting the maximum amount of word replacements to 20% of the input sequence length. We apply the same threshold to the RANDOM and PRIORITIZED attacks, but not to PWWS since we observed low replacement rates despite the attack's effectiveness. This is in agreement to the results reported in Ren et al. (2019).

C Frequency differences for CNN and LSTM models

The $\log_e$ frequencies for the four attacks against the CNN and LSTM models can be found in Table 4. In accordance to the experiments with RoBERTa (see Section 3 in the paper), we observe large Cohen's d effect sizes for the majority of the comparisons, which shows that the statistical frequency differences between replaced words and their substitutions are present for adversarial attacks against these two models as well.

D Bayes factors

The Bayes factors for the mean frequency comparisons between replaced words and their adversarial substitutions can be found in Table 5. We observe high values for BF_{10} across all comparisons, providing strong evidence for the hypothesis that the $\log_e$ frequency means between replaced words and their substitutions are different.

E Visualizations of frequency differences

Figure 3 illustrates the frequency differences for attacks against the RoBERTa model using his-

Dataset	Model	Attack	Subst.	non-OOV
IMDb	CNN	RANDOM	$> 10^{10594}$	$> 10^{7004}$
		PRIORITIZED	$> 10^{6549}$	$> 10^{5009}$
		GENETIC	$> 10^{2581}$	$> 10^{2318}$
		PWWS	$> 10^{2182}$	$> 10^{1673}$
	LSTM	RANDOM	$> 10^{9643}$	$> 10^{6338}$
		PRIORITIZED	$> 10^{5949}$	$> 10^{4967}$
		GENETIC	$> 10^{2550}$	$> 10^{2369}$
		PWWS	$> 10^{1666}$	$> 10^{1442}$
	RoBERTa	RANDOM	$> 10^{12138}$	$> 10^{7948}$
		PRIORITIZED	$> 10^{9014}$	$> 10^{6043}$
		GENETIC	$> 10^{4215}$	$> 10^{3672}$
		PWWS	$> 10^{5182}$	$> 10^{3656}$
SST-2	CNN	RANDOM	$> 10^{754}$	$> 10^{138}$
		PRIORITIZED	$> 10^{573}$	$> 10^{222}$
		GENETIC	$> 10^{388}$	$> 10^{104}$
		PWWS	$> 10^{397}$	$> 10^{131}$
	LSTM	RANDOM	$> 10^{800}$	$> 10^{153}$
		PRIORITIZED	$> 10^{648}$	$> 10^{264}$
		GENETIC	$> 10^{522}$	$> 10^{148}$
		PWWS	$> 10^{456}$	$> 10^{144}$
	RoBERTa	RANDOM	$> 10^{867}$	$> 10^{149}$
		PRIORITIZED	$> 10^{779}$	$> 10^{130}$
		GENETIC	$> 10^{584}$	$> 10^{55}$
		PWWS	$> 10^{600}$	$> 10^{125}$

Table 5: Bayes factors (BF_{10}) for the Bayesian hypothesis tests.

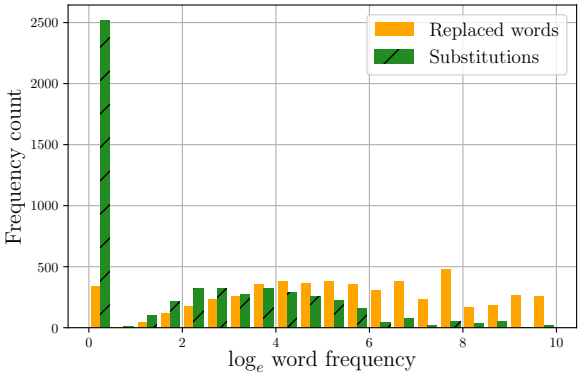
tograms. We observe that for the majority of the attacks, OOV substitutions occur most often among the perturbed sequences.

F Varying false positive thresholds

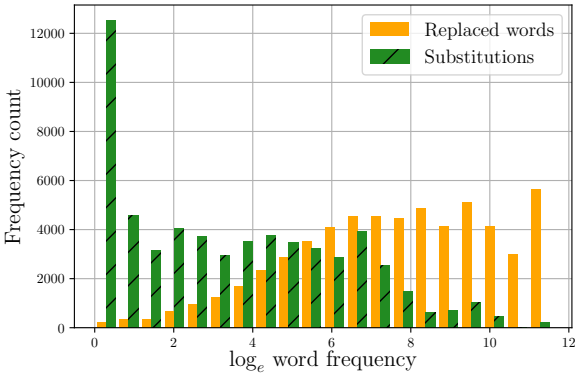
The rate of false positives predicted by a detection system is crucial for its practicability, and a limited amount of false positives is hence highly desirable. Figure 4 illustrates the true positive rates predicted by FGWS for all attacks against RoBERTa with different quasi-fixed false positive thresholds (as in Section 4.2 of the paper, δ was tuned on the validation set for each value of γ corresponding to the specific false positive threshold). As expected, we observe a trade-off between true and false positive rates for varying values of γ, such that lower false positive rates imply lower true positive rates. However, even for false positive rates of 1% and 5%, we observe that FGWS is able to detect between 33.6% and 90.0% of adversarial examples on IMDb and between 31.7% and 67.2% on SST-2. This indicates that FGWS has the potential to detect a useful fraction of adversarial examples without creating an excessive burden of false positives.

G Additional FGWS examples

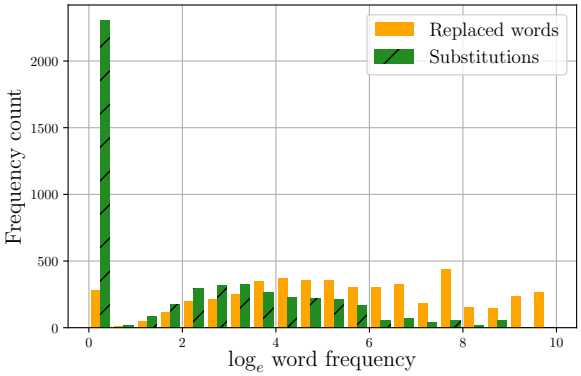
Additional examples of FGWS can be found in Table 6 (true positives), Table 7 (false positives), Table 8 (true negatives) and Table 9 (false negatives).



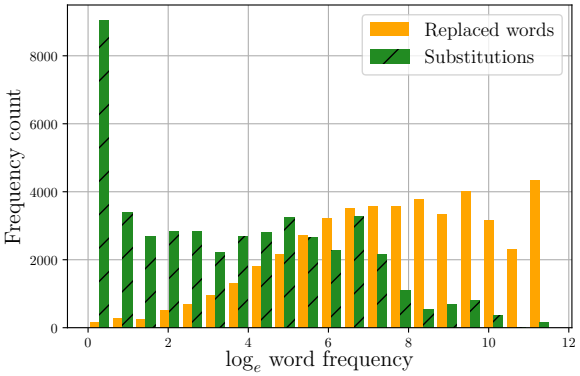
(a) RANDOM on SST-2



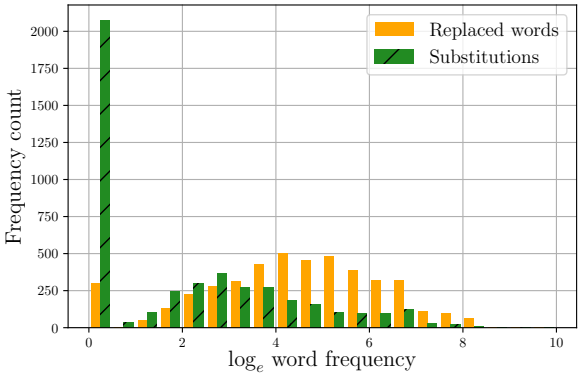
(b) RANDOM on IMDb



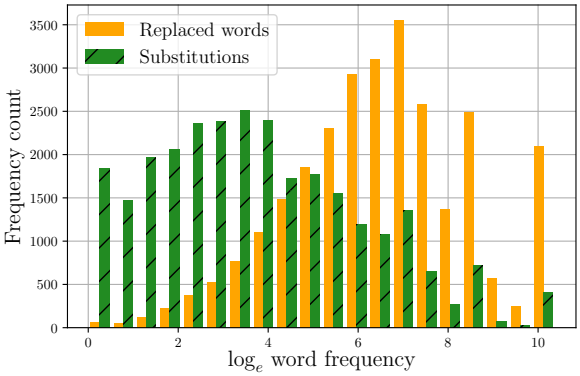
(c) PRIORITIZED on SST-2



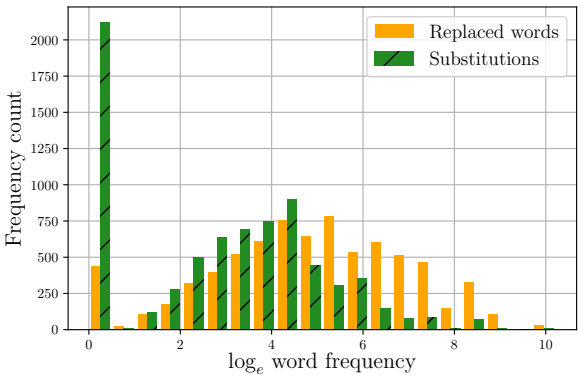
(d) PRIORITIZED on IMDb



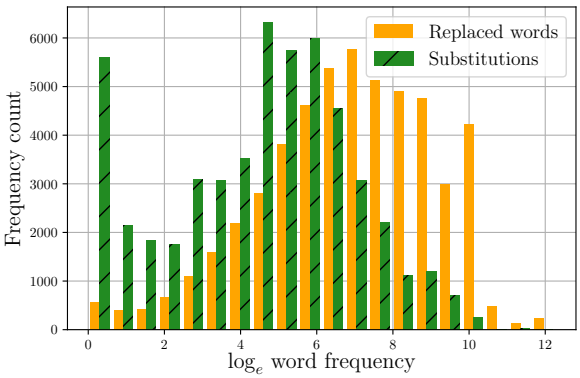
(e) GENETIC on SST-2



(f) GENETIC on IMDb

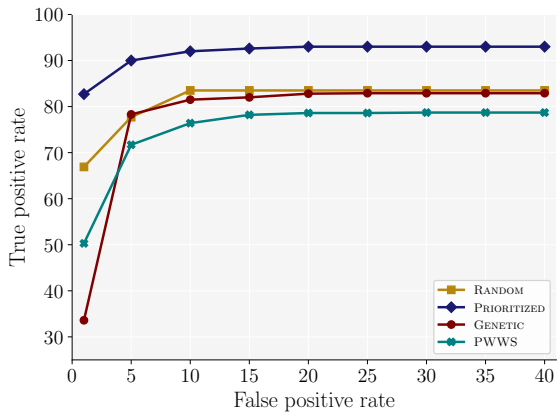


(g) PWWS on SST-2

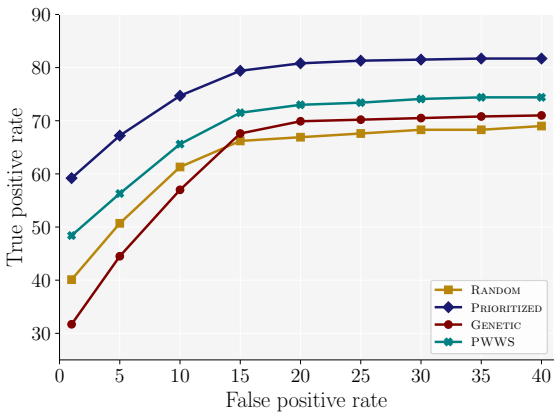


(h) PWWS on IMDb

Figure 3: Histograms showing the frequency distribution of words replaced by the attacks and their corresponding substitutions against the RoBERTa model. The x-axis represents the words' $\log_e$ frequency with respect to the model's training corpus, the y-axis denotes their respective frequencies among the perturbed test set sequences.



(a) IMDb



(b) SST-2

Figure 4: The trade-off between true and false positive rates on the test sets with all four attacks against RoBERTa on (a) IMDb and (b) SST-2. The true positive rates (y-axis) are computed when γ is set to allow for different quasi-fixed amounts of false positives (x-axis).

Model:RoBERTa on SST-2

Unperturbed	first good then bothersome	*negative* (74.5%)
GENETIC	first good then **galling** [***bothersome***]	*positive* (88.7%)
DISP	first good **that** [***then***] galling	*positive* (84.8%)
FGWS	first good then **annoying** [***galling***]	*negative* (91.3%)

Model:RoBERTa on IMDb

Unperturbed	i am a huge rupert everett fan . i adore kathy bates so when i saw it available i decided to check it out . the synopsis didn t really tell you much . in parts it was silly touching and in others some parts were down right hysterical . any person that is a huge fan of a personality of any type will find some small identifying traits with the main character . of course there are many they won t but that is the point if you like any of the actors give it a watch but don t look for any thing too dramatic it s good fun . i might also mention you can see how darn tall rupert is . i mean i knew he was 6 4 but he seems even more in this film . he even seemed to stoop a bit due to the other characters height in this . he is tall i mean tall and for you rupert fans there is a bare chest scene ... wonderful	*positive* (99.2%)
PWWS	i am a huge rupert everett fan . i adore kathy bates so when i saw it available i decided to **stop** [***check***] it out . the synopsis didn t really tell you much . in parts it was silly touching and in others some parts were down right hysterical . any person that is a huge fan of a personality of any type will find some small identifying traits with the main character . of course there are many they won t but that is the point if you like any of the actors give it a watch but don t look for any thing too dramatic it s **undecomposed** [***good***] fun . i might also mention you can see how darn tall rupert is . i mean i knew he was 6 4 but he seems even more in this film . he even seemed to stoop a bit **imputable** [***due***] to the other characters height in this . he is tall i mean tall and for you rupert fans there is a bare chest scene ... **tremendous** [***wonderful***]	*negative* (60.1%)
DISP	i am a huge rupert everett fan . i adore kathy bates so when i saw it available i decided to **out** [***stop***] it out . the synopsis didn t really tell you much . in parts it was silly touching and in others some parts were down right hysterical . any person that is a huge fan of a personality of any type will find some small identifying traits with the main character . of course there are many they won **,** [***t***] but that is the point if you like any of the actors give it a watch but don t look for any thing too dramatic it **'s** [***s***] **so** [***undecomposed***] fun . i **can** [***might***] also mention you can see how darn tall rupert is . i mean i knew he was 6 4 but he seems even more in this film . he even seemed to stoop a bit imputable to the other characters height in this . he is tall i mean tall and for you rupert fans there is a bare chest scene ... **.** [***tremendous***]	*positive* (92.0%)
FGWS	i am a huge rupert everett fan . i adore kathy bates so when i saw it available i decided to stop it out . the synopsis didn t really tell you much . in parts it was silly touching and in others some parts were down right hysterical . any person that is a huge fan of a personality of any type will find some small **place** [***identifying***] traits with the main character . of course there are many they won t but that is the point if you like any of the actors give it a watch but don t look for any thing too dramatic it s **good** [***undecomposed***] fun . i might also mention you can see how darn tall rupert is . i mean i knew he was 6 4 but he seems even more in this film . he even seemed to **sit** [***stoop***] a bit **due** [***imputable***] to the other characters height in this . he is tall i mean tall and for you rupert fans there is a bare chest scene ... tremendous	*positive* (88.9%)

Table 6: Illustration of true positives generated with FGWS against RoBERTa on SST-2 (top) and IMDb (bottom). The substitutions caused the model to change the predicted label back to its ground-truth for the given adversarial examples.

Model: RoBERTa on SST-2

Unperturbed	imagine if you will a tony hawk skating video interspliced with footage from behind enemy lines and set to jersey shore techno	*negative* (83.6%)
DISP	imagine if you $^{5.97}$**get** [$^{6.84}$***will***] a tony hawk skating video $^{0.00}$**,** [$^{0.00}$***interspliced***] with footage from behind enemy lines and set to jersey shore techno	*negative* (87.1%)
FGWS	imagine if you will a $^{3.76}$**kevin** [$^{1.10}$***tony***] $^{3.93}$**pitch** [$^{2.48}$***hawk***] skating video interspliced with footage from behind enemy lines and set to $^{6.55}$**new** [$^{2.08}$***jersey***] $^{3.93}$**sea** [$^{2.08}$***shore***] $^{5.69}$**music** [$^{1.61}$***techno***]	*positive* (65.7%)

Model: RoBERTa on IMDb

Unperturbed	admittedly alex has become a little podgey but they are still for me the greatest rock trio ever . i wholeheartedly recommend this dvd to any fan . i was very disappointed that they canceled their planned recent munich gig logistics and regret not making an effort to see them elsewhere . the dvd is a small consolation the greatest incentive to acquire a proper dvd playback setup . naive perhaps but i still don t understand the significance of the tumble driers on stage i would be grateful for any clarification . cheers iain .	*positive* (99.4%)
DISP	admittedly alex has become a little podgey but they are still for me the greatest rock trio ever . i wholeheartedly recommend this dvd to any fan . i was very disappointed that they canceled their planned recent munich gig logistics and regret not making an effort to see them elsewhere . the dvd is a small consolation the greatest incentive to acquire a proper dvd playback setup . naive perhaps but i still don t understand the significance of the $^{9.77}$**one** [$^{1.95}$***tumble***] driers on stage i would be grateful for any clarification . cheers $^{10.73}$**that** [$^{0.00}$***iain***] .	*positive* (99.3%)
FGWS	admittedly alex has become a little podgey but they are still for me the greatest rock trio ever . i $^{4.55}$**disagree** [$^{2.30}$***wholeheartedly***] recommend this dvd to any fan . i was very disappointed that they canceled their planned recent $^{5.03}$**germany** [$^{2.08}$***munich***] gig $^{3.40}$**transport** [$^{0.69}$***logistics***] and regret not making an effort to see them elsewhere . the dvd is a small $^{5.69}$**win** [$^{2.08}$***consolation***] the greatest $^{5.53}$**opportunity** [$^{1.61}$***incentive***] to acquire a proper dvd $^{6.21}$**editing** [$^{2.08}$***playback***] setup . naive perhaps but i still don t understand the significance of the $^{6.19}$**fall** [$^{1.95}$***tumble***] $^{1.61}$**dryer** [$^{0.00}$***driers***] on stage i would be grateful for any $^{5.11}$**explanation** [$^{1.10}$***clarification***] . cheers iain .	*negative* (50.1%)

Table 7: Illustration of false positives generated with FGWS against RoBERTa on SST-2 (top) and IMDb (bottom). The substitutions caused the model to change the predicted label for the given unperturbed sequences.

Model: RoBERTa on SST-2

Unperturbed	it s a hoot and a half and a great way for the american people to see what a candidate is like when he s not giving the same 15 cent stump speech	*positive* (100.0%)
DISP	it **'s** [*s*] a hoot and a half and a great way for the american people to see what a candidate is like when he **'s** [*s*] not giving the same 15 **minutes** [*cent*] **the** [*stump*] speech	*positive* (100.0%)
FGWS	it s a hoot and a half and a great way for the american people to see what a **nomination** [*candidate*] is like when he s not giving the same 15 cent **stamp** [*stump*] **words** [*speech*]	*positive* (100.0%)

Model: RoBERTa on IMDb

Unperturbed	it was awful plain and simple . what was their message where was the movie going with this it has all the ingredients of a sub b grade movie . from plotless storyline the bad acting to the cheesey slow mo cinematography . i d sooner watch a movie i ve already seen like goodfellas a bronx tale even grease . there are no likeable characters . in the end you just want everyone to die already . save 2 hours of your life and skip this one .	*negative* (99.9%)
DISP	it was awful plain and simple . what was their message where was the movie going with this it has all the ingredients of a sub b grade movie . from plotless storyline the bad acting to the cheesey slow mo cinematography . i **would** [*d*] sooner watch a movie i **have** [*ve*] already seen like goodfellas a bronx tale **in** [*even*] grease . there are no likeable characters . in the end you just want everyone to die already . save 2 hours of your life and skip this one .	*negative* (99.9%)
FGWS	it was awful plain and simple . what was their message where was the movie going with this it has all the ingredients of a sub b grade movie . from **unwatchable** [*plotless*] storyline the bad acting to the **cheesy** [*cheesey*] slow mo cinematography . i d sooner watch a movie i ve already seen like goodfellas a bronx tale even grease . there are no likeable characters . in the end you just want everyone to die already . save 2 hours of your life and skip this one .	*negative* (99.9%)

Table 8: Illustration of true negatives generated with FGWS against RoBERTa on SST-2 (top) and IMDb (bottom). The substitutions did not cause the model to change the predicted label for the given unperturbed sequences.

Model: RoBERTa on SST-2

Unperturbed	the spark of special anime magic here is unmistakable and hard to resist	*positive* (100.0%)
PWWS	the spark of special anime **deception** [***magic***] here is unmistakable and **laborious** [***hard***] to **hold** [***resist***]	*negative* (84.4%)
DISP	the spark of special anime deception here is unmistakable and **able** [***laborious***] to hold	*positive* (99.9%)
FGWS	the spark of special anime deception here is **subtle** [***unmistakable***] and laborious to hold	*negative* (97.8%)

Model: RoBERTa on IMDb

Unperturbed	graduation day is a result of the success of friday the 13th . both of those films are about creative bloody murders rather than suspense . if you enjoy that type of film i d recommend graduation day . if not i wouldn t. there s nothing new here just the same old killings . even though i ve given the film a 4 out of 10 i will say that it s not a repulsive film . it is watchable if your curious about it just not creative .	*negative* (71.3%)
GENETIC	graduation day is a result of the success of friday the 13th . both of those films are about creative bloody murders rather than suspense . if you enjoy that type of film i d recommend graduation day . if not i wouldn t. there s nothing new here just the same **ancient** [***old***] killings . even though i ve given the film a 4 out of 10 i will say that it s not a repulsive film . it is watchable if your curious about it just not creative .	*positive* (53.5%)
DISP	graduation day is a result of the success of friday the 13th . both of those films are about creative bloody murders rather than suspense . if you enjoy that type of film i **would** [***d***] recommend graduation day . if not i **do** [***wouldn***] t. there **is** [***s***] nothing new here just the same ancient killings . even though i **have** [***ve***] given the film a 4 out of 10 i will say that it **'s** [***s***] not a **good** [***repulsive***] film . it is watchable if your curious about it **is** [***just***] not creative .	*negative* (99.5%)
FGWS	graduation day is a result of the success of friday the 13th . both of those films are about creative bloody murders rather than suspense . if you enjoy that type of film i d recommend graduation day . if not i wouldn t. there s nothing new here just the same ancient killings . even though i ve given the film a 4 out of 10 i will say that it s not a repulsive film . it is watchable if your curious about it just not creative .	*positive* (53.5%)

Table 9: Illustration of false negatives generated with FGWS against RoBERTa on SST-2 (top) and IMDb (bottom). The substitutions did not cause the model to change the predicted label back to its ground-truth for the given adversarial examples.