Identifying Human Strategies for Generating Word-Level Adversarial Examples

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

Abstract

Adversarial examples in NLP are receiving increasing research attention. One line of investigation is the generation of word-level adversarial examples against fine-tuned Transformer models that preserve naturalness and grammaticality. Previous work found that human- and machine-generated adversarial examples are comparable in their naturalness and grammatical correctness. Most notably, humans were able to generate adversarial examples much more effortlessly than automated attacks. In this paper, we provide a detailed analysis of exactly how humans create these adversarial examples. By exploring the behavioural patterns of human workers during the generation process, we identify statistically significant tendencies based on which words humans prefer to select for adversarial replacement (e.g., word frequencies, word saliencies, sentiment) as well as *where* and *when* words are replaced in an input sequence. With our findings, we seek to inspire efforts that harness human strategies for more robust NLP models.

1 Adversarial attacks in NLP

Researchers in natural language processing (NLP) have identified the vulnerability of machine learning models to adversarial attacks: controlled, meaning-preserving input perturbations that cause a wrong model prediction (Jia and Liang, 2017; Iyyer et al., 2018; Ribeiro et al., 2018). Such adversarial examples uncover model failure cases and are a major challenge for trustworthiness and reliability. While several defence methods exist against adversarial attacks (Huang et al., 2019; Jia et al., 2019; Zhou et al., 2019; Jones et al., 2020; Le et al., 2022), developing robust NLP models is an open research challenge. An in-depth analysis of word-level adversarial examples, however, identified a range of problems, showing that they are often ungrammatical or semantically inconsistent (Morris et al., 2020).[1] This finding raised the question of how feasible natural and grammatically correct adversarial examples actually are in NLP.

To answer this question, Mozes et al. (2021a) explored whether humans are able to generate adversarial examples that are valid under such strict requirements. In that study, crowdworkers were tasked with the generation of word-level adversarial examples against a target model. The findings showed that at first sight—without strict validation—humans are less successful than automated attacks. However, when adding constraints on the preservation of sentiment, grammaticality and naturalness, human-authored examples do not differ from automated ones. The most striking finding was that automated attacks required massive computational effort while humans were able to do the same task using only a handful of queries.[2] This suggests that humans are far more efficient in adversarial attacks than automated systems, yet exactly how they achieve this is unclear.

In this work, we address this question by analysing human behaviour through the public dataset from Mozes et al. (2021a). We look at which words humans perturbed, where within a sentence those perturbations were located, and whether they mainly focused on perturbing sentiment-loaded words. We find that *(i)* in contrast to automated attacks, humans use more frequent adversarial word substitutions, *(ii)* the semantic similarity between replaced words and adversarial substitutions is greater for humans than for most attacks, and *(iii)* humans replace sentiment-loaded words more often than algorithmic attackers. Our goal is to understand what makes humans so efficient at this task, and whether these strategies could be harnessed for more adversarially robust NLP models.

[1]For example, replacing the word *summer* with *winter*.

[2]For example, 140,000 queries are needed per example for SEMEMEPSO (Zang et al., 2020), on average, to generate successful adversarial examples on IMDb (Maas et al., 2011), whereas humans need 10.9 queries (Mozes et al., 2021a).

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6118–6126
December 7-11, 2022

Attack	All			Successful			Unsuccessful		
	Δ_M	Δ_{SD}	d	Δ_M	Δ_{SD}	d	Δ_M	Δ_{SD}	d
HUMANADV	0.6	3.1	0.2	0.5	3.0	0.1	0.6	3.1	0.2
TEXTFOOLER	2.5	2.6	0.8	2.5	2.6	0.8	2.5	2.6	0.8
GENETIC	1.5	2.1	0.5	1.4	2.0	0.5	1.5	2.1	0.5
BAE	2.0	4.0	0.5	1.9	4.1	0.5	2.0	4.0	0.5
SEMEMEPSO	2.4	2.8	0.8	2.4	2.8	0.8	–	–	–

Table 1: Word frequency differences between replaced words and adversarial substitutions. Δ_M and Δ_{SD} represent the mean and standard deviation of the differences between replaced words and substitutions (i.e., positive values: replaced words > substitutions), d denotes the Cohen's d effect size. Note that for SEMEMEPSO, all adversarial examples are successful.

2 Data and Models

We present a fine-grained analysis of the strategies that human crowdworkers employed to generate word-level adversarial examples against sentiment classification models. In the dataset from Mozes et al. (2021a), 43 participants were recruited via Amazon Mechanical Turk and trained to perform a word-level adversarial attack on test set sequences from the IMDb movie reviews dataset (Maas et al., 2011). In total, 170 adversarial examples were collected. For each of the collected adversarial examples, the authors also generated automated adversarial examples using the TEXTFOOLER (Jin et al., 2019), BAE (Garg and Ramakrishnan, 2020), GENETIC (Alzantot et al., 2018) and SEMEMEPSO (Zang et al., 2020) attacks.

The TEXTFOOLER attack uses a greedy word-replacement algorithm that is guided by word saliencies and semantic similarity measures between an unperturbed sequence and the adversarial candidate. The BAE attack resorts to a different technique, utilising a BERT-based language model to remove and replace tokens in an input sequence. The GENETIC attack, in contrast, is based on a population-based method using genetic algorithms. Finally, the SEMEMEPSO attack is based on replacements of word sememes instead of entire words and combines this with a particle swarm optimisation approach.

All attacks were performed against a RoBERTa model (Liu et al., 2019) fine-tuned on IMDb.[3] Here, we only consider adversarial examples that preserved sentiment after evaluation by an independent set of crowdworkers, which Mozes et al. (2021a) used as a key validity criterion.

[3]For more model details, see Section 3 in Mozes et al. (2021a).

3 Analysis

In this section, we report on a series of experiments analysing the human- and machine-authored adversarial examples.

3.1 What do humans replace?

Word frequency. We investigate the word frequency of the adversarial examples. Existing work (Mozes et al., 2021b; Hauser et al., 2021) identified significant differences in word frequency between adversarially perturbed words (hereafter referred to as *replaced words*) and their substitutions (hereafter referred to as *adversarial substitutions*) for a number of attacks. The substituted words were considerably less frequent than their original counterparts (e.g., *annoying* → *galling*).[4] Here, we examine whether this pattern is also evident in humans' strategies. Table 1 shows the differences of the $\log_e$ word frequencies between replaced words and corresponding substitutions for all four automated adversarial attacks and the human attack. All attacks replace words with less frequent substitutions. The notable observations here are the human-authored examples: the $\log_e$ frequency differences are lowest for the human-generated substitutions (HUMANADV). The effect size Cohen's d, which expresses the absolute magnitude of the effect that frequencies differ, further shows that the high-to-low frequency replacement is much less used by humans ($d = 0.2$) than by the other, automated attacks ($d \geq 0.5$). These findings persist when inspecting either successful or unsuccessful adversarial examples in isolation.

To test for statistical differences between the attacks, we first conduct a 5 (attacks) by 2 (success) ANOVA on the $\log_e$ frequency differences between replaced words and substitutions, to determine whether main effects or interaction effects were present. We observe a significant main effect for attack, $F(4, 12003) = 152.85, p < .001$, but none for success nor an interaction between attack and success.[5]

Overall, the results suggest that humans use a strategy different from automated approaches and find replacements that do not rely on the high-to-low frequency mapping to the same extent as automated attacks. Illustrations of the highest and

[4]Word frequency is computed with respect to the model's training corpus in these experiments.

[5]Follow-up experiments revealed significant differences between HUMANADV and all attacks.

lowest frequency differences among word substitution pairs can be found in the Appendix (Table 4).

Word saliency usage. In the crowdsourcing study by Mozes et al. (2021a), humans were provided with the word saliency information (i.e., individual words were highlighted based on how much the model's prediction confidence would change if they were deleted).[6] This was originally intended to make the task easier for humans. Now, we investigate whether humans did indeed focus on salient words.

Did humans prefer salient words? First we investigate whether the saliency of a replaced word correlates with the iteration index at which this word was selected for replacement by a human crowdworker.[7] Across all examples, we obtain a negligible negative Pearson correlation of $r = -0.05$ ($p < 0.01$). However, the correlation is weak, which does not provide additional evidence in favour of utilising saliencies for automated attacks based on human behaviour.

Did salient words lead to successful attacks? We furthermore analyse whether the average saliency across all replaced words of a sequence correlates with whether this led to a successful (i.e., label-flipping) adversarial example. For each valid human-generated adversarial example, we hence compute the point-biserial correlation between attack success and mean saliency of replaced words. The findings suggest that the higher the saliency of replaced words, the higher the chance of success of an adversarial example, $r = 0.26$ ($p < .006$). Analogously, we also computed the correlation between the mean word saliency across all replaced words per iteration and the corresponding decrease in prediction confidence. The findings indicate a small correlation of $r = 0.12$ ($p < .001$): replacing a more salient word leads to larger increases in prediction confidence change.

It is worth noting that, even though the word saliency is defined as the decrease in prediction confidence after deleting a word from the sequence, this finding is not necessarily expected: a human attacker not only needs to identify and remove an existing word in the sequence, but they also have to find a semantically suitable replacement that decreases the model's prediction confidence.

It is furthermore worth mentioning that both BAE and TEXTFOOLER define the token importance rankings based on a word saliency measure, and therefore explicitly incorporate the word saliency into the attack process. The results obtained in this work provide additional evidence in favour of utilising saliencies for automated attacks, showing that humans (which have been shown to generate adversarial examples in a more efficient way) also tend to utilise word saliencies.

[6] It is worth noting that we cannot be certain whether humans did indeed use the word saliencies during the process.

[7] An iteration index of 1 means that a word was the first to be replaced.

Attack	Valid pairs	All	Succ.	Unsucc.
HUMANADV	990/1303	0.47 (0.19)	0.52 (0.18)	0.44 (0.19)
TEXTFOOLER^{a,b}	1497/1805	0.57 (0.20)	0.58 (0.20)	0.53 (0.16)
GENETIC^{b}	1955/2437	0.44 (0.19)	0.44 (0.18)	0.44 (0.19)
BAE^{a,b}	940/1623	0.69 (0.25)	0.70 (0.24)	0.69 (0.26)
SEMEMEPSO^{b}	724/946	0.66 (0.17)	0.66 (0.17)	–

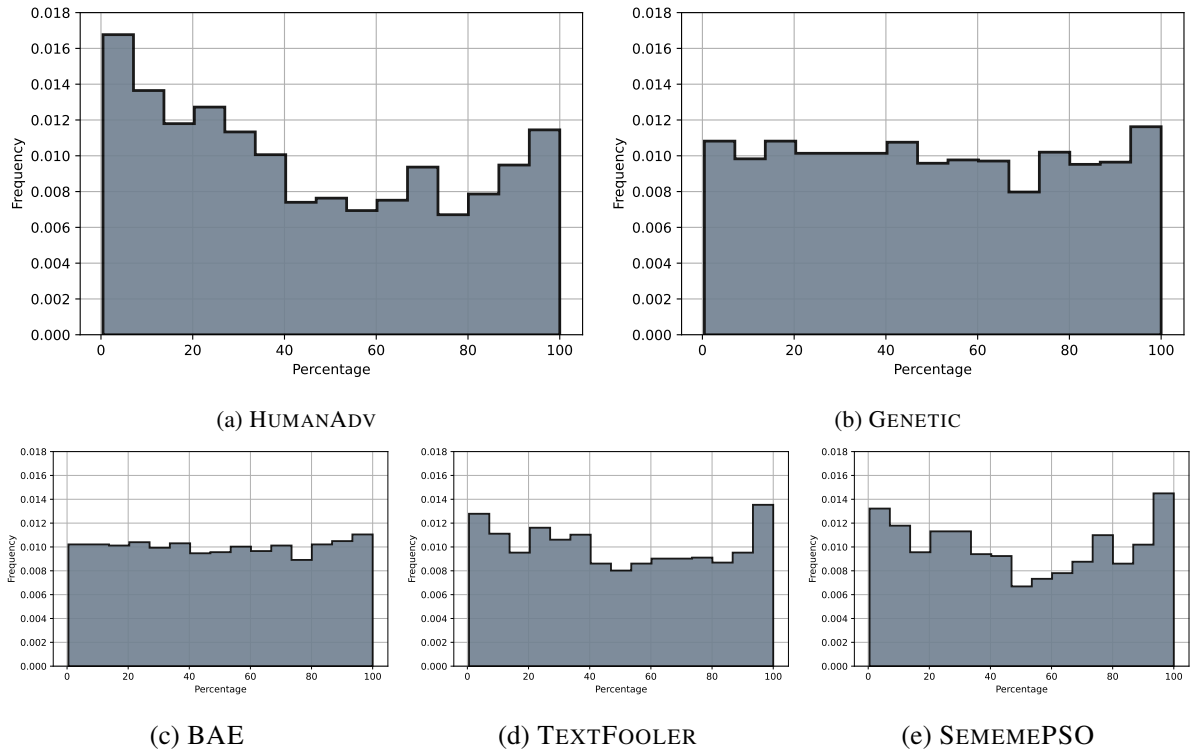
Table 2: The mean (SD) cosine distances between replaced words and substitutions. ^{a} indicates significant differences with HUMANADV for unsuccessful pairs, ^{b} for successful ones.

Attack	All		Succ.		Unsucc.	
	Rep.	Sub.	Rep.	Sub.	Rep.	Sub.
HUMANADV	22.9	20.7	23.7	24.0	22.5	19.3
TEXTFOOLER^{b}	19.8	14.2	19.8	14.3	18.8	12.5
GENETIC^{b}	19.7	14.3	20.3	15.7	19.6	14.0
BAE^{a,b}	16.5	4.3	19.3	5.3	15.8	4.0
SEMEMEPSO	21.8	20.8	21.8	20.8	–	–

Table 3: Ratio (%) of replaced (Rep.) and adversarially substituted words (Sub.) with existing sentiment value. ^{a} indicates significant differences with HUMANADV for replaced words, ^{b} for substitutions.

Word similarities. Next, we compare the semantic differences in adversarial substitution pairs across the different attacks. While the algorithmic attacks source word synonyms from available lexical databases such as WORDNET (Fellbaum, 1998) or GLOVE embeddings (Pennington et al., 2014), humans directly choose word replacements based on their own vocabulary and can therefore use substitutions that more accurately fit the context of the replaced word. Hence, we might expect to see a difference between the semantic similarity of human- and machine-chosen substitutions.

To test this idea, we compare the pre-trained word embeddings for the replaced words and their corresponding substitutions. We choose counter-fitted GLOVE embeddings (Mrkšić et al., 2016), as they push synonyms further together and antonyms further apart in representation space.



(a) HUMANADV (b) GENETIC

(c) BAE (d) TEXTFOOLER (e) SEMEMEPSO

Figure 1: Histograms visualising the distribution of index percentages at which the adversarial attacks perturb individual input words.

Table 2 shows the cosine distances of the embeddings between the pairs for all five attacks. Valid pairs denotes the fraction of valid pairs used to compute the distances, since some of the word pairs did not have embedding representations in the used space. To test for statistical effects, we conduct a 5 (attacks) by 2 (success) ANOVA on the cosine distances between embeddings of replaced words and corresponding substitutions, revealing significant main effects for attack, $F(4, 6097) = 363.63, p < .001$, success, $F(1, 6097) = 16.43, p < .001$, as well as an interaction effect, $F(3, 6097) = 5.54, p < .001$. The entangled significant differences between attacks are indicated in Table 2. For success, a t-test reveals significant differences ($p < .001$) between successful and unsuccessful cosine distances across attacks. For their interaction, the difference could be driven by the lack of observations given for the unsuccessful SEMEMEPSO pairs.

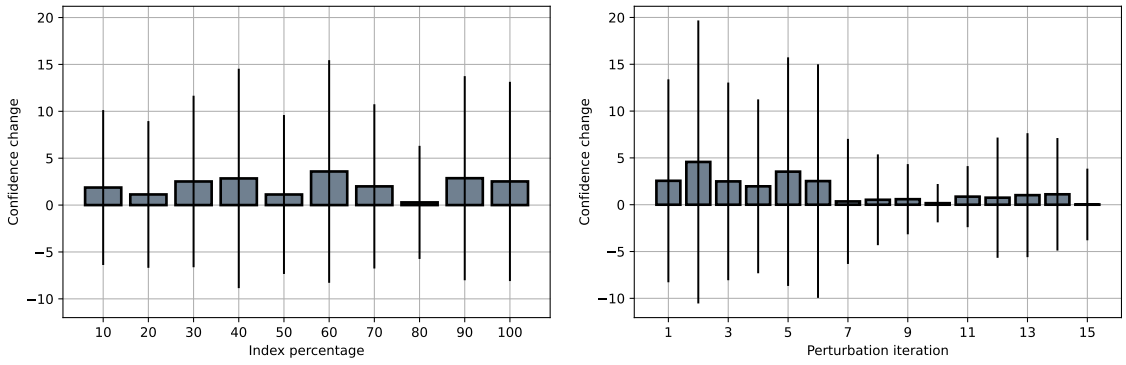
The findings indicate that human-generated adversarial substitution pairs are significantly more similar than the substitution pairs of automated attacks (all except GENETIC). A possible explanation for this variability is that GENETIC uses counter-fitted embedding spaces for identifying semantically-related words for adversarial substitution. However, TEXTFOOLER uses the same embedding representations, yet the distances appear to be larger. Illustrative examples of semantically similar and dissimilar word substitution pairs can be found in the Appendix (Table 5).

Repeating the analysis with regular GLOVE embeddings yields similar results, albeit without an interaction effect (see Appendix A). We furthermore provide an analysis of sentence similarities between adversarial examples in Appendix B.

3.2 How many replaced words have sentiment value?

Particularly for the task of sentiment analysis, an attack might be more successful if it focuses on words with a sentiment value (e.g., *like*, *great*). We investigate the differences between attacks with respect to how many replaced words and adversarial substitutions have sentiment value. To do this, we compute the ratio of replaced words (to all replaced input words) that have a sentiment value in the NLTK sentiment lexicon (Loper and Bird, 2002). Table 3 reveals that this sentiment ratio is low (between 16% and 23%) across attacks.

For replaced words, we observe a significant



(a) (b)

Figure 2: Mean (standard deviation) prediction confidence changes on the true class across examples with respect to (a) the word index percentage and (b) the iteration in which human crowdworkers change individual input words.

main effect for attack, $F(4, 8105) = 5.28, p < .001$, but not for success or their interaction. For adversarial substitutions, the same ANOVA yields a significant effect for attack, $F(4, 8105) = 54.64, p < .001$, but likewise not for success or their interaction. HUMANADV and SEMEMEPSO tend to follow that strategy more so than the remaining attacks.[8] We provide illustrations of the substitution pairs with the highest increases and decreases in sentiment in the Appendix (Table 8).

3.3 Where do humans replace?

Next, we investigate the specific regions in an input sequence (e.g., start, middle, end) where adversarial attacks prefer to perturb words. To do this, we define the index percentage of a word in an input sequence as the ratio of the word's index to the number of words in the input (e.g., the third word of a sequence of ten words would have an index percentage of 30%).

Figure 1 shows the frequency of index percentages per attack and suggests that HUMANADV, TEXTFOOLER and SEMEMEPSO preferentially perturb words at the beginning and end of an input sequence. In contrast, the distributions for BAE and GENETIC show a uniform pattern. For GENETIC this result is somewhat expected: the attack selects words for replacement by sampling words proportionately to their number of available synonyms rather than based on a semantically-informed strategy. The HUMANADV's preference for replacing words at the beginning and the end of the sequence could be explained by the attention that humans devote to these parts of the text when reading from left to right. Perhaps most interestingly, the distributions for TEXTFOOLER and BAE differ, despite both using word saliencies as their word importance ranking.

We investigate which individual word changes led to notable changes in prediction confidence of the target model. We first analyse this by looking at the relationship between the index percentage and the change in prediction confidence on the true class (Figure 2). We observe that (a) the confidence changes caused by human perturbations are not prevalent at a specific index percentage, but rather distribute fairly evenly across the start, middle and end of the sequence. Second, Figure 2 (b) shows that the confidence changes are higher in the first iterations, and seem to drastically reduce after the sixth iteration on average.

4 Discussion and conclusion

This work presented a granular analysis on strategies followed by humans when attempting to generate adversarial examples through word-level substitutions. We have shown that the difference in word frequency between replaced words and adversarial substitutions is smaller for humans than for the automated attacks. Furthermore, humans tend to use substitutions that are more semantically similar to the replaced words than most attacks, and humans target words that have a sentiment value to a larger extent than automated attacks. Based on the findings provided, future directions could focus on harnessing such strategies to improve existing adversarial attacks and in doing so ultimately increase the robustness of machine learning-based NLP models against adversarial attacks.

[8] This observation could potentially be explained by the finding that humans tend to over-perceive word saliencies for words with a strong sentiment value (Schuff et al., 2022).

Ethical considerations

This paper discusses adversarial attacks in NLP, methods that are developed to uncover failure cases of machine learning models, and specifically potential approaches to further enhance such attacks against text classification models. It is worth mentioning that these methods can be used maliciously, for example to circumvent content filtering systems for hateful or offensive language on social media. Our work is intended to better understand the phenomenon of adversarial examples in NLP, its relation to human language understanding, and to harness such insights to contribute to more robust models against adversarial input perturbations.

Limitations

The presented work comes with a number of limitations which will be discussed in this section.

First, our analyses are limited to a single target dataset (IMDb movie reviews) and based on the only existing "human word-level adversarial attacks" dataset. Replicating our experiments on other datasets, especially those containing different styles of language use such as formal academic or journalistic writing, would help to further understand the behavioural patterns used by humans when generating adversarial examples. Future work could also build on the approach of Mozes et al. (2021a) to collect a larger dataset that would allow us to learn more about the strategies employed by humans when crafting adversarial examples.

Second, additional linguistic and behavioural patterns could potentially be analysed in the data. We primarily focused on the central aspects driving human strategies, yet there are other dimensions on which the data can be inspected for additional behavioural patterns (e.g., part-of-speech usage by human attackers). These are beyond the scope of this contribution but could in the future inform better attack and defence models.

Third, the dataset from Mozes et al. (2021a) did not contain potential moderating variables about the human crowdworkers. As a consequence, it is unknown how or whether differences in, for example, the language proficiency of participants, experience with NLP crowdsourcing tasks or even general cognitive abilities played a role. While the authors applied some participation requirements (i.e., participation in a similar NLP study) and trained the crowdworkers, the next step would be to understand whether psychological variables potentially moderate one's ability to craft valid adversarial examples.

Finally, the analyses in this work solely focus on statistical data analysis and do not harness data-driven machine learning-based methods to identify behavioural patterns in the data. Nevertheless, in this context the dataset size (170 human-generated sequences) represents a limitation and is potentially not large enough in size to be useful for learning-based experiments. Future work with larger datasets would mitigate that limitation and possibly help generate more insights about human strategies in adversarial example generation.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Jens Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. *arXiv preprint arXiv:2109.07403*.

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 4083–4093, 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 4129–4142, 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*.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2752–2765, Online. Association for Computational Linguistics.

Thai Le, Noseong Park, and Dongwon Lee. 2022. SHIELD: Defending textual neural networks against multiple black-box adversarial attacks with stochastic multi-expert patcher. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6661–6674, Dublin, Ireland. Association for Computational Linguistics.

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*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

Maximilian Mozes, Max Bartolo, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021a. Contrasting human-and machine-generated word-level adversarial examples for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8258–8270.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021b. Frequency-guided word substitutions for detecting textual adversarial examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 171–186.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

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.

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.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. *arXiv preprint arXiv:2201.11569*.

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.

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 4904–4913, Hong Kong, China. Association for Computational Linguistics.

Difference	HUMANADV			TEXTFOOLER			GENETIC			BAE			SEMEMEPSO		
High	*bad*	→	*,*	*be*	→	*sont*	*one*	→	*uno*	*tobanga*	→	*i*	*movie*	→	*conga*
	annoying	→	*.*	*like*	→	*iove*	*cast*	→	*foundry*	*challen*	→	*s*	*movie*	→	*cancan*
	of	→	*buttery*	*good*	→	*buen*	*action*	→	*measurements*	*hansika*	→	*s*	*really*	→	*sheerly*
	i	→	*i'am*	*very*	→	*vitally*	*time*	→	*timeframe*	*modulates*	→	*was*	*film*	→	*photoshoot*
	this,	→	*this*	*story*	→	*escudos*	*like*	→	*adores*	*bahrani*	→	*t*	*bad*	→	*hardhearted*
Low	*educational*	→	*teaching*	*frostbite*	→	*frostbitten*	*counselors*	→	*advisors*	*turns*	→	*works*	*appearance.the*	→	*present.the*
	makers	→	*producers*	*movie.*	→	*flick.*	*wrought*	→	*fabricated*	*producers*	→	*makers*	*liked*	→	*supposed*
	very	→	*more*	*years.i*	→	*year.i*	*humour*	→	*mood*	*low*	→	*top*	*manages*	→	*attempts*
	bad	→	*great*	*rajasthan*	→	*bihar*	*nearly*	→	*near*	*match*	→	*co*	*promote*	→	*cheer*
	sing	→	*scream*	*supposed*	→	*felt*	*dirty*	→	*nasty*	*dead*	→	*line*	*died*	→	*failed*

Table 4: The top five pairs of replaced words and adversarial substitutions with the highest and lowest absolute frequency differences across attacks. Pairs were pre-filtered such that at least one word in a pair has a positive frequency in the training corpus, to avoid low differences due to both words having a frequency of zero.

Distance	HUMANADV			TEXTFOOLER			GENETIC			BAE			SEMEMEPSO		
High	*in*	→	*unoriginal*	*like*	→	*iove*	*blood*	→	*chrissakes*	*earlier*	→	*inger*	*movies*	→	*jitterbugs*
	adder	→	*enough*	*story*	→	*escudos*	*brett*	→	*broadly*	*end*	→	*oja*	*box*	→	*flagellation*
	back	→	*askance*	*door*	→	*fatma*	*x*	→	*tenth*	*played*	→	*dermott*	*movie*	→	*cancan*
	guard	→	*kilter*	*link*	→	*nol*	*volunteers*	→	*boneheads*	*guess*	→	*eses*	*series*	→	*wisps*
	jeepers	→	*like*	*camera*	→	*salas*	*barbara*	→	*barbaric*	*put*	→	*udge*	*episode*	→	*triviality*
Low	*could*	→	*would*	*eight*	→	*six*	*would*	→	*could*	*films*	→	*film*	*usually*	→	*generally*
	awful	→	*terrible*	*two*	→	*three*	*become*	→	*becoming*	*dancing*	→	*dance*	*ridiculous*	→	*laughable*
	could	→	*might*	*awful*	→	*terrible*	*awful*	→	*terrible*	*know*	→	*tell*	*positive*	→	*negative*
	anything	→	*something*	*test*	→	*tests*	*cards*	→	*card*	*sort*	→	*kind*	*specific*	→	*particular*
	films	→	*film*	*so*	→	*too*	*investment*	→	*investments*	*unless*	→	*if*	*even*	→	*however*

Table 5: The top five pairs of replaced words and adversarial substitutions with the highest and lowest word embedding cosine distance across attacks (using the counter-fitted embeddings).

Attack	Valid pairs	All	Succ.	Unsucc.
HUMANADV	1109/1303	0.46 (0.21)	0.49 (0.21)	0.45 (0.21)
TEXTFOOLER	1542/1805	0.56 (0.20)	0.57 (0.20)	0.52 (0.17)
GENETIC	2020/2437	0.44 (0.19)	0.45 (0.19)	0.44 (0.19)
BAE	1319/1623	0.71 (0.30)	0.73 (0.29)	0.71 (0.31)
SEMEMEPSO	787/946	0.64 (0.18)	0.64 (0.18)	–

Table 6: The mean (and standard deviation) cosine distances (GLOVE embeddings) between replaced words and corresponding substitutions for the five attacks across all perturbed sequences, divided into all, as well as successful and unsuccessful sequences.

Attack	All	Succ.	Unsucc.
HUMANADV	0.035 (0.050)	0.043 (0.061)	0.031 (0.042)
TEXTFOOLER	0.064 (0.065)	0.063 (0.064)	0.177 (0.000)
GENETIC^a	0.063 (0.052)	0.034 (0.036)	0.076 (0.053)
BAE^a	0.044 (0.036)	0.022 (0.018)	0.056 (0.039)
SEMEMEPSO	0.056 (0.071)	0.056 (0.071)	–

Table 7: The mean (SD) cosine distances of USE representations between unperturbed and adversarial sequences. ^{a} indicates significant differences with HUMANADV for unsuccessful pairs.

A Word similarities

We repeat the experiments in Section 3 for word similarities with regular GLOVE embeddings, rather than the counter-fitted ones. The mean (standard deviation) distances can be found in Table 6. We here also conduct a 5 (attacks) by 2 (success) ANOVA, yielding significant effects for attack, $F(4, 6768) = 371.37, p < .001$, and success, $F(1, 6768) = 11.27, p < .001$, but not for their interaction. To disentangle this effect for success, a subsequent test on an aggregation of successful and unsuccessful word pairs across attacks reveals significant differences ($p < .001$) between both samples. Comparing HUMANADV to all other attacks, we observe statistically significant ($p < .01$) differences between all comparisons for the successful portion of the data. For the unsuccessful ones, only the comparison between HUMANADV and BAE yields significant differences.

B Sentence similarities

Word similarities may only provide a limited picture as they lack context. We therefore also analyse the sentence similarity among adversarial examples. We utilise *universal sentence encoder* (USE, Cer et al., 2018) representations for our analysis. Table 7 shows the cosine distances for each attack type. Conducting a 5 (attack) by 2 (success) ANOVA, we observe significant effects between attacks, $F(4, 627) = 6.46, p < .001$, success, $F(1, 627) = 16.41, p < .001$ as well as their interaction, $F(3, 627) = 5.77, p < .001$.[9]

[9]The results of subsequent t-tests are indicated in Table 7.

Sentiment increase	HUMANADV			TEXTFOOLER			GENETIC			BAE			SEMEMEPSO		
Smallest	*best*	→	*worst*	*comedic*	→	*travesty*	*comedy*	→	*travesty*	*enjoyed*	→	*cut*	*positive*	→	*negative*
	love	→	*hate*	*comedy*	→	*ridicule*	*excited*	→	*agitated*	*reaches*	→	*lies*	*amazing*	→	*horrid*
	enjoyed	→	*hated*	*funny*	→	*odd*	*intense*	→	*violent*	*great*	→	*good*	*great*	→	*terrible*
	excellent	→	*horrible*	*comedy*	→	*farce*	*enlightening*	→	*sobering*	*brilliant*	→	*worthy*	*amazing*	→	*terrible*
	fantastic	→	*bad*	*wonderful*	→	*funky*	*kiss*	→	*screwing*	*fantastic*	→	*good*	*wonderfully*	→	*suspiciously*
Largest	*worst*	→	*best*	*worst*	→	*greatest*	*odd*	→	*curious*	*bad*	→	*good*	*awful*	→	*awesome*
	bad	→	*great*	*worse*	→	*greatest*	*strangely*	→	*surprisingly*	*ridiculous*	→	*good*	*terrible*	→	*terrific*
	idiotic	→	*excellent*	*annoys*	→	*excites*	*cruel*	→	*ferocious*	*dead*	→	*hard*	*awful*	→	*terrific*
	poor	→	*great*	*disappointments*	→	*excitements*	*fine*	→	*beautiful*	*low*	→	*top*	*awful*	→	*thrilling*
	fail	→	*excellent*	*dullest*	→	*neatest*	*worst*	→	*gravest*	*worth*	→	*worthy*	*hard*	→	*great*

Table 8: The top five pairs of replaced words and adversarial substitutions with the largest increases and decreases in sentiment value across attacks (based on the NLTK sentiment lexicon).