Contrasting Human- and Machine-Generated Word-Level Adversarial Examples for Text Classification

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

Abstract

Research shows that natural language processing models are generally considered to be vulnerable to adversarial attacks; but recent work has drawn attention to the issue of validating these adversarial inputs against certain criteria (e.g., the preservation of semantics and grammaticality). Enforcing constraints to uphold such criteria may render attacks unsuccessful, raising the question of whether valid attacks are actually feasible. In this work, we investigate this through the lens of human language ability. We report on crowdsourcing studies in which we task humans with iteratively modifying words in an input text, while receiving immediate model feedback, with the aim of causing a sentiment classification model to misclassify the example. Our findings suggest that humans are capable of generating a substantial amount of adversarial examples using semantics-preserving word substitutions. We analyze how human-generated adversarial examples compare to the recently proposed TEXTFOOLER, GENETIC, BAE and SEMEMEPSO attack algorithms on the dimensions naturalness, preservation of sentiment, grammaticality and substitution rate. Our findings suggest that human-generated adversarial examples are not more able than the best algorithms to generate natural-reading, sentiment-preserving examples, though they do so by being much more computationally efficient.

1 Introduction

The vulnerability of natural language processing (NLP) models to adversarial examples has received widespread attention (Alzantot et al., 2018; Iyyer et al., 2018; Ren et al., 2019). Text processing models have been shown to be susceptible to adversarial input perturbations across tasks, including question answering and text classification (Jia and Liang, 2017; Jin et al., 2019). The concept of adversarial examples originated in computer vision (Szegedy et al., 2013; Goodfellow et al., 2014), and in that domain defines perturbations of input data to neural networks that are barely perceptible to the human viewer. Due to the discrete nature of text, however, that definition is less applicable in an NLP context, since every perturbation to the input tokens is unavoidably perceptible. Consequently, recent work aims to perturb textual inputs while preserving the sequence's naturalness and semantics (i.e., rendering changes imperceptible on these dimensions). However, as shown by Morris et al. (2020a), achieving these desiderata is challenging because even small perturbations can render a text meaningless, grammatically incorrect or unnatural, and furthermore several proposed adversarial attacks fail routinely to achieve them. If the algorithms are modified to ensure that they do achieve the desiderata then their rate of generating successful examples greatly diminishes, suggesting that the reported success rates of recently proposed attacks might represent an overestimation of their true capabilities. This, in turn, raises the question of whether valid word-level adversarial examples are routinely possible against trained NLP models.

In this work, we aim to address this question by incorporating human judgments into the adversarial example generation process. Specifically, we report on a series of data collection efforts in which we task humans to generate adversarial examples from existing movie reviews, while instructed to strictly adhere to a set of validity constraints. In contrast to previous work (e.g., Bartolo et al., 2020; Potts et al., 2020), and in an attempt to replicate a word-level attack's mode of operation, human participants were only able to substitute individual words, and were not allowed to delete or insert new words into the sequence. This represents a black-box attack scenario, since human participants do not have access to information about the model's parameters or gradients. Participants worked in a web interface (Figure 1) that allowed them to conduct word-level substitutions while receiving

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8258–8270
November 7–11, 2021.

immediate feedback from a trained model.

After collecting the human-generated adversarial examples, we compare them to a set of automated adversarial examples for the same sequences using four recently proposed attacks: TEXTFOOLER (Jin et al., 2019), GENETIC (Alzantot et al., 2018), BAE (Garg and Ramakrishnan, 2020), and SEMEMEPSO (Zang et al., 2020). Using human judgments from an independent set of crowdworkers, we assess for each generated adversarial example (human and automated) whether the perturbations changed the sequence's overall sentiment and whether they remained natural.

We find that humans are capable of generating label-flipping word-level adversarial examples (i.e., the classifier misclassifies the sequence after human perturbation) in approximately 50% of the cases. However, when comparing the ground truth labels of perturbed sequences to the sentiment labels provided by the independent set of human annotators, we find that only 58% of the label-flipping human adversarial examples preserve their target sentiment after perturbation. This is considerably lower than for the best automated attacks, which exhibit a label consistency of up to 93% (TEXTFOOLER) after perturbation. In terms of naturalness, we find no statistically significant differences between the human and machine attacks for the majority of comparisons. We furthermore observe that the human-generated sequences introduce fewer grammatical errors than most attacks.

These findings show that under similar constraints, machine-generated, word-level adversarial examples are comparable to human-generated ones with respect to their naturalness and grammaticality. Importantly, however, humans require, on average, only 10.9 queries to run against the model to generate label-flipping adversarial examples, while some attacks require thousands. We believe that our findings could further push the development of reliable word-level adversarial attacks in NLP, and our method and data might aid researchers in identifying human-inspired, more efficient ways of conducting adversarial word substitutions against neural text classification models.

The remainder of this paper is structured as follows. Section 2 discusses previous work related to our research. Section 3 describes both phases of our data collection approach, i.e., the human generation of word-level adversarial examples and the subsequent validation of human- and machine-generated sequences with respect to their preservation of semantics and naturalness. This is followed by the analysis reported in Section 4, and a discussion of our findings and future work in Section 5. Finally, we conclude our paper in Section 6.

2 Related work

Adversarial attacks for NLP. Adversarial attacks have been increasingly applied to NLP, with a diverse set of attack types being investigated, ranging from character-level edits (Ebrahimi et al., 2018b,a), word-level replacements (Alzantot et al., 2018), adding text to the input (Jia and Liang, 2017), paraphrase-level modifications (Iyyer et al., 2018; Ribeiro et al., 2018), to creating adversarial examples from scratch (Bartolo et al., 2020; Nie et al., 2019). This work focuses on word-level attacks.

Word-level attacks. Our work builds on existing efforts on word-level adversarial attacks. Attacks of this type can be further distinguished by whether the adversary has access to the model parameters (i.e., white-box) or is restricted to accessing only the predicted labels or confidence scores (i.e., black-box) (Yuan et al., 2019). Word-level attacks have been explored for NLP tasks such as question answering (Blohm et al., 2018; Welbl et al., 2020), natural language inference (Jin et al., 2019), and text classification (Papernot et al., 2016; Jin et al., 2019). A range of methodologies has been explored for finding optimal synonym substitutions, including population-based gradient-free optimization via genetic algorithms (Alzantot et al., 2018), word saliency probability weighting (Ren et al., 2019), similarity and consistency filtering (Jin et al., 2019), sememe-based word substitution and particle swarm optimization-based search (Zang et al., 2020), and contextual perturbations from masked language models (Garg and Ramakrishnan, 2020). Word-level perturbations have also been used as part of data augmentation strategies to certifiably improve model robustness (Jia et al., 2019).

Existing efforts to detect and defend against word-level adversarial examples resort to adversarial data augmentation (e.g., Ren et al., 2019; Jin et al., 2019) as well as rule-based (Mozes et al., 2021) and learning-based (Zhou et al., 2019) approaches to identify adversarially perturbed inputs.

Evaluating word-level attacks. Of particular importance for this paper is how adversarial at-

tacks can be evaluated against various dimensions. Adversarial attack performance has been shown to vary across evaluation dimensions including adversarial success rates, readability and content preservation (Xu et al., 2020), as well as linguistic constraints such as semantics, grammaticality, the edit distance between original and perturbed text, and non-suspicion (Morris et al., 2020a). However, to the best of our knowledge, such evaluation efforts are limited to automated attacks and how humans perform at creating word-level adversarial examples across these dimensions remains unexplored.

Human-in-the-loop adversarial examples. When the task is unconstrained, human crowdworkers have been shown to be capable of creating high quality adversarial examples for a variety of NLP tasks such as question answering (Wallace et al., 2019; Dua et al., 2019; Bartolo et al., 2020; Khashabi et al., 2020), natural language inference, (Nie et al., 2019), and sentiment analysis (Potts et al., 2020). We extend this line of work and investigate whether human capabilities for creating adversarial examples persist when the examples are constrained to arise from word-level perturbations, which have been shown to be highly effective (Alzantot et al., 2018; Jin et al., 2019).

3 Method

Our data collection process has two stages: first, we ask human annotators to perform a word-level adversarial attack for given input sequences. To this end, we prepared an online interface that lets participants perturb input sequences on a word-level whilst receiving immediate feedback as to how their changes affected classifier confidence. Second, we ask an independent set of crowdworkers to evaluate the generated adversarial examples.

3.1 Stage one: human-generated word-level adversarial examples

In order to familiarize participants with the concept of word-level adversarial attacks for stage one of the data collection, we lead them through a sequence of four subtasks, each building on the preceding one:

1. Participants are asked to freely write a movie review with a specified sentiment
2. Participants are asked to freely write an adversarial example
3. Participants are given an existing movie review and are asked to use word-level adversarial perturbations *without* adhering to semantic preservation and grammatical correctness
4. Same as 3, but with the constraints to preserve semantics and grammatical correctness

The data collected in tasks 1, 2 and 3 are not further analyzed in this paper, since these tasks were intended to help participants understand adversarial examples for text classification. After having successfully completed the three preparation tasks, the participants are considered fit to conduct task 4, which is the main topic of interest in this paper. For each subtask, we ask participants to submit four instances. For tasks 3 and 4, we randomly select four test set samples from the IMDb movie reviews dataset (Maas et al., 2011) for each participant. The reference classifier is a RoBERTa model (Liu et al., 2019) fine-tuned on IMDb, as it has been shown to perform highly on this task.[1] Our fine-tuned model achieves an accuracy of 93.8% on the IMDb test set.[2]

For tasks 1 and 2, the participants were able to directly see the classifier prediction before they submitted their reviews through clicking a button that queries the current sequence against the sentiment classification model. For tasks 3 and 4, we asked participants to submit at least 15 iterations of word-level substitutions before moving on to the next review.[3] After each submitted iteration the model provided immediate feedback as to how the change affected its prediction. The sequence of display of the four reviews in tasks 3 and 4 is based on the review length in ascending order.

Word saliencies. For tasks 3 and 4, the interface additionally displays the word saliencies (Li et al., 2016a,b) for each word in the movie review. Here, the word saliency is defined as the model's difference in prediction confidence before and after replacing the word with an out-of-vocabulary token. The interface for tasks 3 and 4 is shown in Figure 1.[4]

[1] Specifically, we use a RoBERTA-base model provided by `HuggingFace` (Wolf et al., 2019), with 125 million parameters.

[2] We randomly sample 1,000 training set sequences for epoch validation, and the final selected model achieves an accuracy of 92.7% on this validation set.

[3] We tested the task with different numbers of iterations, and found this number to be suitable for our experiments.

[4] Participants were given the option to disable the word saliency highlighting, and were also able to undo and redo

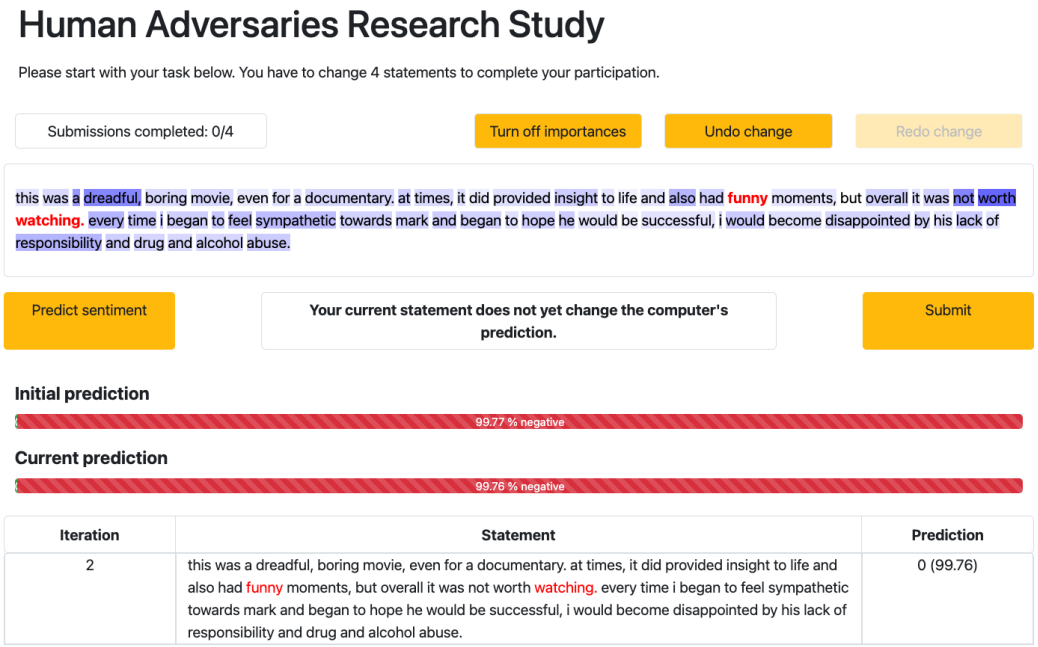


Figure 1: The interface for tasks 3 and 4. Participants are asked to change individual words in existing movie reviews to lead the RoBERTa model into misclassification. The word color highlighting represents the respective saliencies for each word in the sequence (see Section 3.1 for details).

We use Amazon's Mechanical Turk to collect the data. We restrict participation to workers that have previously conducted more than 1,000 successful Human Intelligence Tasks (HITs), have an approval rate of above 98% and who are located in Canada, the US, or the UK. We estimate the completion time to be under 60 minutes, and pay USD 12.40 per user per HIT. In total, we collected responses from $n = 43$ participants. For task 4, we had to exclude two individual submissions due to technical errors. The resulting sample consists of 172 collected reviews for the first three tasks and 170 reviews for task 4. Despite a random allocation of test set sequences to participants, we did not encounter duplicate sequences in the sample.[5]

Comparison to automated attacks. We compare the human-generated, word-level adversarial examples against a set of automatically generated ones. Specifically, we attack the fine-tuned RoBERTa model as used for the data collection phase on the 170 sequences collected in task 4. We experiment with four recently proposed attacks.

GENETIC. The GENETIC attack (Alzantot et al., 2018) uses a population-based genetic search method to generate word-level adversarial examples. Specifically, the attack iteratively adds individual perturbations to an input sequence until the model misclassifies the perturbed input.

TEXTFOOLER. TEXTFOOLER (Jin et al., 2019) is a black-box word-level adversarial attack that ranks words according to their importance for classifier decision-making, and then iteratively replaces the selected words with semantically similar ones to lead the model into misclassification. TEXTFOOLER ensures that the replacement tokens have the same part-of-speech as the selected word. Furthermore, the algorithm utilizes the Universal Sentence Encoder (Cer et al., 2018) to identify replacements that best preserve sequence semantics.

SEMEMEPSO. Whereas existing work predominantly relies on embedding spaces or thesauri like WordNet (Fellbaum, 1998), Zang et al. (2020) propose an attack using sememes (which the authors describe as minimum semantic units of language) to identify semantics-preserving word substitutions. The attack, referred to as SEMEMEPSO, additionally uses a combinatorial optimization method based on particle swarm optimization.

BAE. In contrast to previous approaches, Garg and Ramakrishnan (2020) propose BERT-based Adversarial Examples (BAE), an attack that relies on a BERT masked language model used to

changes made to the input sequence.

[5]The data are available at http://github.com/maximilianmozes/human_adversaries.

both replace and insert new tokens into an existing sequence to generate an adversarial example. They introduce multiple variants of BAE and in this work, we experiment with the BAE-R variant, which only replaces tokens, but does not insert new ones. This is to ensure that BAE is directly comparable to the other attacks analyzed in our experiments.

We generate adversarial examples based on the 170 sequences used during the data collection study, and use the `TextAttack` (Morris et al., 2020b) Python library with all attacks in their default configuration. For computational efficiency, for the GENETIC attack, we use a slightly different variant compared to Alzantot et al. (2018). Specifically, we use the `faster-alzantot` variant offered by `TextAttack`, which implements the modifications suggested in Jia et al. (2019).

3.2 Stage two: evaluating generated adversarial examples

To evaluate the adversarial examples generated by algorithmic approaches and human participants in stage one, we ask an independent set of crowdworkers to annotate the collected data. Specifically, in a new data collection stage, participants read and judged each adversarial example on its sentiment and naturalness, both on a five-point Likert scale. Here, a rating of 1 would denote very negative sentiment (a very unnatural review), whereas a rating of 5 would indicate a very positive sentiment (a very natural review). We use the sentiment judgments to measure the deviation of sentiment resulting from introducing the perturbations (high deviations imply a larger shift in sentiment), and the naturalness judgment to evaluate whether the adversarial substitutions distort the naturalness of the sequence. Specifically, we ask participants to rate the 172 generated adversarial examples from task 2, the 170 unperturbed reviews used in task 4, and the corresponding human- and machine-generated adversarial examples. For the examples in task 4, we select the first label-flipping iteration for a successful submission, and the iteration which exhibits the lowest confidence on the ground truth for unsuccessful submissions.

We recruited participants via the Prolific Academic[6] platform, and aimed to collect three independent ratings per text. We used independent workers per criterion and recruited 120 participants for each. Each participant was asked to rate 30 texts (randomly selected from all available sequences) and received GBP 1.50 as compensation. On average, each text was rated by 3.55 human judges.

Attack	ASR	Reference	`TextAttack`
HUMAN	48.8	—	—
GENETIC	38.2	42.9	46.7
TEXTFOOLER	99.4	98.8	100.0
BAE	43.0	42.3	55.6
SEMEMEPSO	100.0	100.0	100.0

Table 1: Attack success rates (ASR) on the 170 test set sequences. Reference denotes the success rate against an independent fine-tuned RoBERTa model, `TextAttack` refers to the success rates reported by Morris et al. (2020b) against a BERT-Base model using 100 random sequences from IMDb.

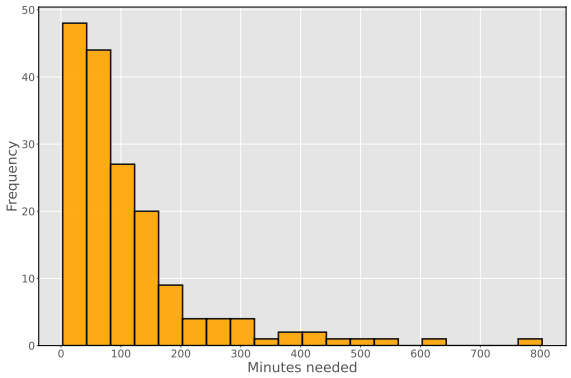


Figure 2: Minutes needed by participants for task 4.

4 Analysis

After collecting the human judgments we analyze both the human and machine attacks' performance on generating adversarial examples. The primary objective is to investigate the feasibility of word-level adversarial examples that adhere to validity criteria as suggested in previous work (Morris et al., 2020a). We use the *attack success rate* (ASR) as the initial metric to evaluate the performance of either attack mode (human and algorithmic). The attack success rate is defined as the percentage of successful adversarial examples (i.e., those that are misclassified after perturbation) to all perturbed sequences.

We observe that overall, workers were generally able to generate successful movie reviews for task 1 (for 90% of the submitted sequences the model predicted the desired sentiment) and led the model into misclassification in task 2 for the majority of the cases (ASR 80%). For task 3, workers also

[6] https://www.prolific.co/

Attack	Match (S)	Δ_S	Match (U)	Δ_U
HUMAN	58%	1.15 (1.10)	90%	0.35 (0.82)
GENETIC	86%	0.33 (0.85)	98%	0.23 (0.65)
TEXTFOOLER	93%	0.28 (0.68)	100%	0.60 (0.00)
BAE	82%	0.29 (0.88)	97%	0.29 (0.52)
SEMEMEPSO	82%	0.47 (0.89)	—	—

Table 2: The percentage of sentiment-preserving adversarial examples per attack. Match (S) denotes the percentage of label-flipping (successful) samples that preserve sentiment, Match (U) denotes unsuccessful ones.

managed to flip the model prediction by introducing arbitrary word-level perturbations (ASR 86%). Crucially, when we introduced constraints in task 4, the ASR drops to 49%, suggesting an increased difficulty of generating word-level adversarial examples when attempting to preserve the sentiment and naturalness of the text. It is worth mentioning that we conducted additional experiments with expert annotators (i.e., academic researchers with experience in NLP) and found that the ASR for task 4 was even lower compared to the crowdworkers. As a comparison, we report the ASR of all word-level attacks in Table 1, and observe that the HUMAN ASR is higher than the ones for GENETIC and BAE, but lower than TEXTFOOLER and SEMEMEPSO.

Figure 2 depicts the distribution of times needed for the human participants to generated the word-level adversarial examples in task 4. We observe that participants need on average 111.29 minutes (standard deviation: 119.77) to complete the task.

4.1 Analysis of human annotations

Sentiment. We define the final sentiment value for each text as negative if its mean rating is below 3.0, and positive if above.[7] As an initial test, we compute the correlation between the ground truth label (positive or negative) and the mean human sentiment rating for unperturbed samples for task 4. We obtain a Pearson correlation of $r = 0.89$ (95% CI $= [0.85, 0.92]$, $p < .001$). This demonstrates high agreement between the IMDb ground truth labels and the human annotations for both tasks.

Next, we want to assess whether adversarial examples preserve the sentiment of the original sequence. To test this, we compare the ground truth label for each text with its binarized human sentiment label and consider sentiment to have been preserved when these agree. Table 2 shows the proportion of adversarial examples whose ground truth label matches the binarized human rating. Δ_S and Δ_U represent the mean (standard deviation) differences in ratings between the original and adversarial sequences. The higher the difference, the more do human ratings between the unperturbed and perturbed sequences deviate from each other.

All algorithmic attacks show high values (above 80%) for successful examples, while the HUMAN attacks preserve the sentiment less often (58%). Similarly, the mean distance ($\Delta_S = 1.15$) for the HUMAN attack is considerably higher than that for the algorithmic attacks. Thus, of the human-generated adversarial examples, only 58% preserve the original sentiment and can be considered for further evaluation. The central question now is whether the higher sentiment-preservation rate of algorithmic attacks holds up if we submit the data to a naturalness test.

Naturalness. Similar to sentiment, we now compare the naturalness ratings between the unperturbed and attacked sequences. The average naturalness rating per text is compared between unperturbed texts and their adversarial counterparts. The larger that difference, the more unnatural the adversarial perturbations have rendered the respective movie review. We only consider the sentiment-preserving adversarial examples as explained in Section 4.1.

To test statistically, whether the attacks differed in their naturalness deviation, we ran a 5 (*attack types*) by 2 (*success:* successful and unsuccessful) ANOVA with the naturalness differences as the dependent variable. That analysis yielded a significant main effect of attack type, $F(4, 666) = 7.87, p < 0.001$ and success, $F(1, 666) = 18.64, p < 0.001$, both of which were subsumed in the interaction effect, $F(3, 666) = 7.29, p < 0.001$.

To disentangle the interaction effect, we show the Cohen's d effect sizes (Cohen, 1988) for the attack type comparisons for successful and unsuccessful attacks. This analysis helps us to understand how the effect of attack type depends on the attack's success. The effect size d expresses the absolute magnitude of the mean naturalness difference per comparison and is preferred over p-values.[8] Table 3 shows the d values with their

[7] 80 samples with a mean rating of exactly 3.0 were excluded from our analysis.

[8] $d = 0.2$, $d = 0.5$ and $d = 0.8$ can be interpreted as a small, medium and large effects, respectively.

	HUMAN	GENETIC	BAE	TEXTFOOLER	SEMEMEPSO
HUMAN	—	-0.32 $[-0.79; 0.16]$	-0.98 $[-1.49; -0.47]$*	—	—
GENETIC	-0.55 $[-1.22; 0.12]$	—	-0.63 $[-1.09; -0.17]$*	—	—
BAE	-0.23 $[-0.88; 0.42]$	0.29 $[-0.33; 0.91]$	—	—	—
TEXTFOOLER	0.13 $[-0.41; 0.67]$	0.65 $[0.13; 1.16]$*	0.35 $[-0.14; 0.85]$	—	—
SEMEMEPSO	-0.26 $[-0.81; 0.30]$	0.27 $[-0.25; 0.79]$	-0.02 $[-0.53; 0.48]$	-0.38 $[-0.76; -0.01]$*	—

Table 3: Cohen's d effect sizes for naturalness comparisons. The lower triangle represents comparisons for successful adversarial examples, the upper one those for unsuccessful examples. The table can be read row-wise, such that the rating differences are computed by subtracting the mean of the column attack from the mean of the row attack (i.e., a negative effect size indicates that the mean naturalness difference of the row attack is lower than that of the column attack). * denotes statistically significant differences.

Attack	Δ_S	Δ_U	Δ_{comb}
HUMAN	0.50 (1.25)	0.14 (1.33)	0.27 (1.31)
GENETIC	-0.16 (1.16)	0.55 (1.29)	0.32 (1.29)
TEXTFOOLER	0.67 (1.32)	2.67 (0.00)	0.68 (1.33)
BAE	0.20 (1.33)	1.30 (1.05)	0.89 (1.27)
SEMEMEPSO	0.17 (1.28)	—	0.17 (1.28)

Table 4: The differences (mean and standard deviation) between the average naturalness rating for the unperturbed and attacked sequences for successful (Δ_S) and unsuccessful (Δ_U) adversarial examples as well as their combination (Δ_{comb}). Positive values indicate a decrease in naturalness. Histograms highlighting the distribution of mean ratings can be found in Figure 3 of the Appendix.

99.75% ($p = 0.05/20$) confidence intervals (CI). A CI containing zero implies that the difference in naturalness cannot be considered statistically significant and therefore be disregarded. For the unsuccessful examples, the comparisons are missing for the TEXTFOOLER and SEMEMEPSO attacks. This is because both attacks are highly successful, such that only a single (TEXTFOOLER) and none (SEMEMEPSO) of the adversarial examples did not flip the classifier's prediction.

No differences emerge between the mean naturalness rating difference for the majority of comparisons with respect to the HUMAN attack. Only for the unsuccessful adversarial examples do we see that the rating differences between HUMAN and BAE are significantly different. As a whole, this analysis suggests that in terms of naturalness, the HUMAN adversarial examples are not significantly different from the machine-generated ones (see Table 4 for the means).

4.2 Substitution rate and number of queries

Next, we analyze the effect of the substitution rate for each adversarial example on its corresponding naturalness rating as well as the number of model queries required per attack. Statistical testing with an ANOVA showed that there were significant main effects of attack type and success as well a significant interaction. Table 5 indicates significant differences between the comparisons. Further, we observe a negative Pearson correlation of $r = -0.31$ (95% CI $= [-0.38, -0.24]$, $p < .001$) between the mean naturalness ratings and the word substitution rate, indicating that the naturalness deteriorated with increasing substitutions. Moreover, Table 5 shows that the automated attacks perform notably more model queries as compared to the HUMAN attack.[9] While some attacks query a model thousands of times for a single adversarial example, humans are able to find successful adversarial examples with an average of 10.9 queries run against a model. This suggests that humans are considerably more efficient in generating valid word-level adversarial examples. Together, these findings raise the question of how automated attacks might be further optimized with respect to their computational efficiency.

4.3 Grammaticality

As a last evaluation dimension, we look at the number of grammatical mistakes made between the original reviews and their adversarial counterparts. We follow Morris et al. (2020a) by using the `LanguageTool`[10] grammar checker but exclude all errors related to the category `CASING` since all sequences have been lower-cased. We compare the mean number of grammatical errors made per attack and the percentage of unperturbed-adversarial sequence pairs for which the adversarial example has more grammatical errors than the unperturbed

[9] Note that we do not consider the model queries used for computing the word saliencies provided to the crowdworkers in this comparison.

[10] https://github.com/jxmorris12/language_tool_python

Attack	Sub_S	Sub_U	Q_S	Q_U
HUMAN	7.5 (9.2)	8.6 (8.9)^{a}	10.9 (13.8)	17.5 (10.7)
GENETIC	6.9 (4.2)^{d}	14.0 (4.8)^{c,d}	3558.1 (2102.5)	8069.1 (1211.4)
TEXTFOOLER	8.4 (8.0)^{d,e}	40.3 (0.0)	515.2 (379.3)	1821.0 (0.0)
BAE	4.0 (2.9)^{a,b}	9.6 (1.4)^{a}	292.8 (112.3)	435.8 (149.4)
SEMEMEPSO	5.4 (4.1)^{b}	—	140956.3 (148494.5)	—

Table 5: Mean (SD) substitution rates (Sub) and the number of queries (Q) per attack on all sentiment-preserving adversarial examples. Subscripts S and U denote label-flipping and unsuccessful attacks, respectively. Superscripts indicate significant differences with ^{a}GENETIC, ^{b}TEXTFOOLER, ^{c}HUMAN, ^{d}BAE, and ^{e}SEMEMEPSO attacks.

Attack	Num. errors	Adv. errors (%)
None	10.8 (5.7)*	—
HUMAN	11.2 (5.6)*	34.7
GENETIC	11.1 (5.7)*	37.1
TEXTFOOLER	11.7 (5.7)*	56.5
BAE	15.0 (6.1)*	92.4
SEMEMEPSO	11.0 (5.8)*	22.4

Table 6: Mean (SD) number of errors made per attack and the percentage of cases in which the adversarial example contains more grammatical errors than its unperturbed counterpart (Adv. errors). None represents the unperturbed reviews. *indicates significant difference with BAE.

sequence. For the former, we conduct an ANOVA and compute effect sizes analogously to aforementioned experiments.

Table 6 suggests that all attacks produce texts with a higher number of grammatical errors than the unperturbed sequences. Among the different attacks, BAE generates considerably more grammatical errors (15.0 errors per review) than the other attacks (between 11.0 and 11.7 errors per review). The SEMEMEPSO attack has the lowest rate (22.4%) of increasing grammatical errors. For 34.7% of all tested sequences, the HUMAN adversarial word substitutions yielded an increase in grammatical errors. The percentages of 37.1% for the GENETIC and 56.5% for TEXTFOOLER are comparable to the results reported in Morris et al. (2020a).

Table 7 shows an example movie review from IMDb as well as the perturbed counterparts resulting from all attacks.

5 Discussion

Despite some reported successes, recent work questions the validity of machine-generated word-level adversarial examples. Central to that critical view are evaluation criteria on which the adversarial examples fall short (Morris et al., 2020a). The argument is that with these criteria as constraints, most (if not all) word-level adversarial examples are deemed invalid. In this work, we investigated how feasible such adversarial examples can be generated by humans when explicitly asked to respect a set of validity constraints. The underlying reasoning was that human performance might have been able to improve the quality standard of word-level adversarial examples.

Our findings suggest that with respect to the success rate as well as the preservation of semantics and naturalness, humans do not outperform state-of-the-art attack algorithms in generating word-level adversarial substitutions. But they also do not differ much. This finding speaks to the difficulty of the task. However, our findings suggest that while humans do not outperform machines with respect to the aforementioned criteria, they are able to generate adversarial examples of similar quality using a fraction of the attack iterations required by the algorithms. Humans are able to generate label-flipping examples with only a handful of queries, while the algorithmic attacks might need thousands of inference steps to find successful word substitutions. Further, humans do this without introducing more grammatical errors than the algorithmic attacks. In sum, this work suggests that humans produce adversarial examples comparable to state-of-the-art attacks but at a fraction of the computational costs. With a better understanding of how humans achieve this, future work could try to close that gap and develop more computationally efficient algorithmic adversarial attacks inspired by human language reasoning.

5.1 Limitations and future work

Our work comes with various limitations. First, the broad distribution of human naturalness ratings of unperturbed IMDb test set sequences reflects the in-

Attack	Text	Pred.	Naturalness	Sentiment
—	it boggles the mind how big name stars such as those in this movie can be part of the one of the dullest movies i ve ever seen.	*negative*	4.5	1.9
HUMAN	it **amazes** the mind how big name stars such as those in this movie can be part of the one of the **simplest** movies i ve ever seen.	*positive*	4.3	1.4
GENETIC	it boggles the mind how big **naming** stars such as those in this movie can be part of the one of the dullest **cinema** i ve **always observed.**	*negative*	1.5	1.8
BAE	it boggles the mind how big name stars such as those in this movie can be part of the one of the **liest** movies i ve ever seen.	*positive*	3.7	1.0
TEXTFOOLER	it boggles the mind how big name stars such as those in this movie can be part of the one of the **neatest** movies i ve ever seen.	*positive*	4.0	1.0
SEMEMEPSO	it boggles the mind how big name stars such as those in this movie can be part of the one of the **deepest** movies i ve ever seen.	*positive*	4.3	1.0

Table 7: An example movie review from IMDb together with its corresponding adversarial examples. The Naturalness and Sentiment columns denote the mean ratings as explained in Section 4.1. Individual examples have been reduced to excerpts for better readability, the full texts can be found in Table 8 of the Appendix.

formal style of these texts. Future work would need to assess whether our results would differ in more formal writing (e.g., journalistic or academic writing) where finding adequate replacements while meeting the quality criteria could be even harder. Second, with respect to the number of queries, a direct comparison between the success rates of human and algorithmic attacks might be misleading, since asking humans to conduct thousands of iterations per sequence is practically infeasible. Future work could assess how algorithmic attacks perform if constrained to the same number of iterations as humans.

Moreover, the notable difference in efficiency between humans and algorithms needs to be investigated further, for example by analyzing human strategies in conducting word substitutions, which can potentially be beneficial for developing more efficient attack algorithms.

Additionally, our findings support previous work (Morris et al., 2020a) and suggest that word-level adversarial attacks might impose unrealistic constraints (even on humans). This observation raises the question of whether an attention shift towards phrase-based adversarial examples is needed to guarantee the validity of adversarial examples in NLP. To this end, it would be interesting to expand our research focus beyond word-level attacks, for example by relaxing the constraint on word-level substitutions for humans and giving them additional degrees of freedom to rephrase sequences in individual iterations.

6 Conclusion

This paper compared human and machine performance on generating word-level adversarial examples against a text classification model for sentiment analysis. We observe that human-generated adversarial examples do not preserve a sequence's sentiment as well as machine-generated ones do, but are similar in terms of their naturalness after label-flipping perturbation. While these findings do not suggest that humans outperform algorithms for this task, we find that they achieve similar performance in a much more efficient manner. We therefore believe that our work can build the foundation for future research aiming to further optimize algorithmic word-level attacks by potentially adapting human-inspired strategies for this task.

Acknowledgements

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

Ethical considerations

This work uses publicly available data (Maas et al., 2011) and data collected from human participants. All human participants provided informed consent and the studies were approved by the local ethics review board. No personal information was collected.

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.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. 2018. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 108–118, Brussels, Belgium. Association for Computational Linguistics.

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

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Ebrahimi, Daniel Lowd, and D. Dou. 2018a. On adversarial examples for character-level neural machine translation. In *COLING*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. 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.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

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*.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

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.

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

2020, pages 3829–3839, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020b. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis Griffin. 2021. 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, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pages 49–54. IEEE.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. Dynasent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Johannes Welbl, Pasquale Minervini, Max Bartolo, Pontus Stenetorp, and Sebastian Riedel. 2020. Undersensitivity in neural reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1152–1165, Online. Association for Computational Linguistics.

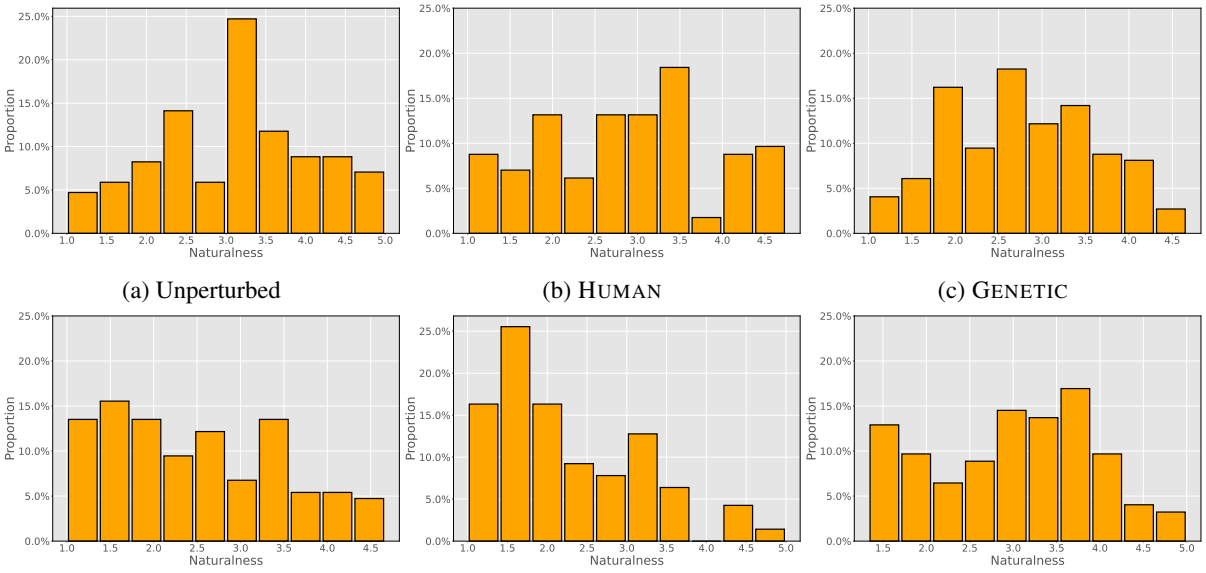
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, and Jey Han Lau. 2020. Elephant in the room: An evaluation framework for assessing adversarial examples in nlp. *arXiv preprint arXiv:2001.07820*.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824.

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.



(a) Unperturbed (b) HUMAN (c) GENETIC

(d) TEXTFOOLER (e) BAE (f) SEMEMEPSO

Figure 3: Histograms of the distribution of mean naturalness ratings across examples for each task (1 = very unnatural, 5 = very natural). For all attacks, only the matched adversarial examples (i.e., those that have an agreement between the annotators' and ground truth sentiment label) were considered.

Attack	Text	Pred.	Naturalness	Sentiment
—	if you are having trouble sleeping or just want to take that nap in the afternoon but just can t seem to drift off, pop in this movie. the only neat thing about this movie are the electric planes. aside from that prepare for some sweet zzzzz s. it boggles the mind how big name stars such as those in this movie can be part of the one of the dullest movies i ve ever seen. now, if you will excuse me, i will finish my nap.	*negative*	4.5	1.9
HUMAN	if you are having **difficulty resting** or just want to take that **break** in the afternoon but just can t seem to drift off, pop in this movie. the only **clever** thing about this movie are the electric planes. aside from that prepare for some **delightful** zzzzz s. it **amazes** the mind how big name stars such as those in this movie can be part of the one of the **simplest** movies i ve ever seen. now, if you will excuse me, i will finish my nap.	*positive*	4.3	1.4
GENETIC	if you are having trouble **asleep** or just **wish** to take that **naps** in the afternoon but just can t seem to drift off, **dad** in this movie. the only **groovy** thing about this **film** are the **electricity airplanes.** aside from that prepare for some sweet zzzzz s. it boggles the mind how big **naming** stars such as those in this movie can be part of the one of the dullest **cinema** i ve **always observed.** now, if you will excuse me, i will **complete** my **naps.**	*negative*	1.5	1.8
BAE	if you are having trouble sleeping or just want to take that nap in the afternoon but just can t seem to drift off, pop in this movie. the only neat thing about this movie are the electric planes. aside from that prepare for some sweet zzzzz s. it boggles the mind how big name stars such as those in this movie can be part of the one of the **liest** movies i ve ever seen. now, if you will excuse me, i will finish my nap.	*positive*	3.7	1.0
TEXTFOOLER	if you are having trouble sleeping or just want to take that nap in the afternoon but just can t seem to drift off, pop in this movie. the only neat thing about this movie are the electric planes. aside from that prepare for some sweet zzzzz s. it boggles the mind how big name stars such as those in this movie can be part of the one of the **neatest** movies i ve ever seen. now, if you will excuse me, i will finish my nap.	*positive*	4.0	1.0
SEMEMEPSO	if you are having trouble sleeping or just want to take that nap in the afternoon but just can t seem to drift off, pop in this movie. the only neat thing about this movie are the electric planes. aside from that prepare for some sweet zzzzz s. it boggles the mind how big name stars such as those in this movie can be part of the one of the **deepest** movies i ve ever seen. now, if you will excuse me, i will finish my nap.	*positive*	4.3	1.0

Table 8: An example movie review from IMDb together with its corresponding adversarial examples.