A Appendix

A.1 Experimental Setup

We ran our experiments on machines running CentOS 7 with 4 TITAN X Pascal GPUs. We fine-tune models on 'BERT-base', which has 110 million parameters. For all datasets, we used the standard train and test split, with the exception of MR, which comes as a single dataset. As is custom, we split the MR dataset into 90% training data and 10% testing data. The samples we chose for each dataset are available in our Github repository along with the results of Mechanical Turk surveys.

A.2 Details about Human Studies.

Our experiments relied on labor crowd-sourced from Amazon Mechanical Turk. We used five datasets: MIT and Yelp datasets from (Alzantot et al., 2018) and MIT, Yelp, and Movie Review datasets from (Jin et al., 2019). We limited our worker pool to workers in the United States, Canada, Canada, and Australia that had completed over 5,000 HITs with over a 99% success rate. We had an additional Qualification that prevented workers who had submitted too many labels in previous tasks from fulfilling too many of our HITs. In the future, will also use a small qualifier task to select workers who are good at the task.

For the human portions, we randomly select 100 successful examples for each combination of attack method and dataset, then use Amazon's Mechanical Turk to gather 10 answers for each example. For the automatic portions of the case study in Section 4, we use all successfully perturbed examples.

A.2.1 Evaluating Adversarial Examples

Rating Semantic Similarity. In one task, we present results from two Mechanical Turk questionnaires to judge semantic similarity or dissimilarity. For each task, we show x and x_{adv}, side by side, in a random order. We added a custom bit of Javascript to highlight character differences between the two sequences. We provided the following description: "Compare two short pieces of English text and determine if they mean different things or the same." We then prompted labelers: "The changes between these two passages preserve the original meaning." We paid $0.06 per label for this task.

Inter-Annotator Agreement. For each semantic similarity prompt, we gathered annotations from 10 different judges. Recall that each selection was one of 5 different options ranging from "Strongly Agree" to "Strongly Disagree." For each pair of original and perturbed sequences, we calculated the number of judges who chose the most frequent option. For example, if 7 choose "Strongly Agree" and 3 chose "Agree," the number of judges who chose the most frequent option is 7. We found that for the examples studied in Section 4 the average of this metric was 5.09. For the examples in Section 5 at the threshold of .98 which we chose, the average was 5.6.

Guessing Real vs. Computer-altered. We present results from our Mechanical Turk survey where we asked users "Is this text real or computer-altered?". We restricted this task to a single dataset, Movie Review. We chose Movie Review because it had an average sample length of 20 words, much shorter than Yelp or IMDB. We made this restriction because of the time-consuming nature of classifying long samples as Real or Fake. We paid $0.05 per label for this task.

Rating word similarity. We performed a third study where we asked showed users a pair of words and asked "In general, replacing the first word with the second preserves the meaning of a sentence:". We paid $0.02 per label for this task.

Phrasing matters. Mechanical Turk comes with a set of pre-designed questionnaire interfaces. These include one titled "Semantic Similarity" which asks users to rate a pair of sentences on a scale from "Not Similar At All" to "Highly Similar." Examples generated by synonym attacks benefit from this question formulation because humans tend to rate two sentences that share many words as "Similar" due to their small morphological distance, even if they have different meanings.

Notes for future surveys . In the future, we would also try to filter out bad labels by mixing some number of ground-truth "easy" data points into our dataset and rejecting the work of labelers who performed poorly on this set.

A.2.2 Finding The Right Thresholds

Comparing two words. We showed study participants a pair of words and asked them whether swapping out one word for the other would change the meaning of a sentence. The results are shown in Figure 3. Using this information, we chose **0.9** as the word-level cosine similarity threshold.

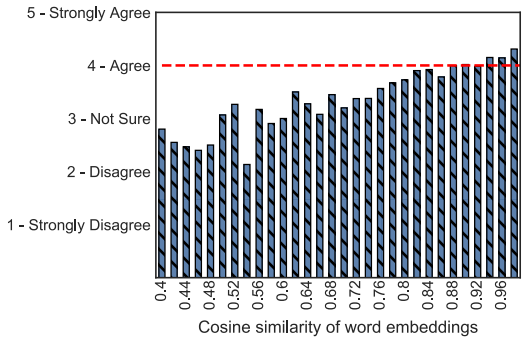

Figure 3: Average response to "In general, replacing the first word with the second preserves the meaning of a sentence" vs. cosine similarity between word1 and word2 (words are grouped by cosine similarity into bins of size .02).

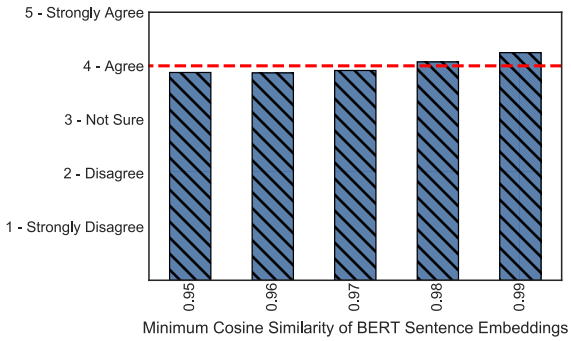

Figure 4: Average response to "The changes between these two passages preserve the original meaning" at each threshold. Threshold is minimum cosine similarity between BERT sentence embeddings.

		Guessed Label	
		Real	Computer-altered
True	Original	814	186
	Perturbed	430	570

Table 8: Confusion matrix for humans guessing if perturbed examples are computer-altered

Comparing two passages. With the word-level threshold set at 0.9, we generated examples at sentence encoder thresholds $\{0.95, 0.96, 0.97, 0.98, 0.99\}$. We chose to encode sentences with a pre-trained BERT sentence encoder fine-tuned for semantic similarity: first on the AllNLI dataset, then on the STS benchmark training set (Reimers and Gurevych, 2019). We repeated the study from 4.1.1 on 100 examples from each threshold, obtaining 10 human labels per example. The results are in Figure 4. On average, judges agreed that the examples produced at **0.98** threshold preserved semantics.

A.3 Further Analysis of Non-Suspicious Constraint Case Study

Table 8 presents the confusion matrix of results from the survey. Interestingly, workers guessed that the examples were real 62.2% of the time, but when they guessed that examples were computer-altered they were right 75.4% of the time. Thus while some perturbed examples are non-suspicious, there are some which workers identify with high precision.

A.4 Adversarial Training Robustness Results

We used our examples for adversarially training by attacking the full MR training set and retraining a new model with the successful examples appended to the training set. Previously, Jin et al. (2019) reported an increase in robustness from adversarial training, while (Alzantot et al., 2018) reported no effect. We trained 5 models on each dataset, and saw significant variance in the robustness of adversarially trained models between random initializations and between epochs. The results are shown in Figure 5. It is possible that Jin et al. (2019) trained a single model for each training set (original and augmented) and happened to see an increase in robustness. It remains to be seen whether examples generated by GENETICATTACK, TEXTFOOLER, and TFADJUSTED help or hurt the robustness and accuracy of adversarially trained models across other model architectures and datasets.

A.5 Word Embeddings

It is common to perform synonym substitution by replacing a word by a neighbor in the counter-fitted embedding space. The distance between word embeddings is frequently measured using Euclidean distance, but it is also common to compare word embeddings based on their cosine similarity (the cosine of the angle between them). (Some work also measures distance based on the mean-squared error between embeddings, which is just the square of Euclidean distance.)

For this reason, past work has sometimes constrained nearest neighbors based on the Euclidean distance between two word vectors, and other times based on their cosine similarity. Alzantot et al. (2018) considered both distance metrics, and reported that they "did not see a noticeable improvement using cosine."

We would like to point out that, when using normalized word vectors (as is typical for

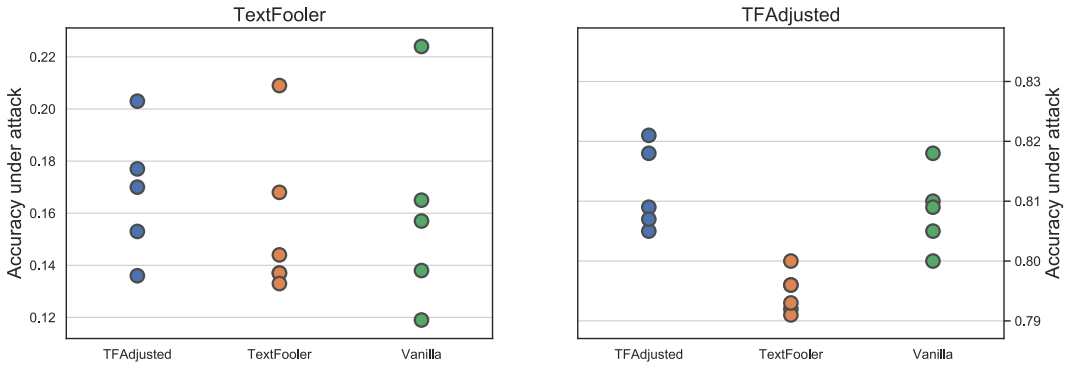

Figure 5: After-attack accuracy of our 15 adversarially trained models subject to two different attacks on the MR test set.

counter-fitted embeddings), filtering nearest neighbors based on their minimum cosine similarity **is equivalent to filtering by maximum Euclidean distance** (or MSE, for that matter).

Proof. Let u, v be normalized word embedding vectors. That is, $\|u\| = \|v\| = 1$. Then $u \cdot v = \|u\|\|v\| \cos(\theta) = \cos(\theta)$.

$$\begin{aligned}\|u - v\|^2 &= (u - v) \cdot (u - v) \\ &= \|u\|^2 - 2(u \cdot v) + \|v\|^2 \\ &= 2 - 2(u \cdot v) \\ &= 2 - 2\cos(\theta).\end{aligned}$$

□

Therefore, the Euclidean distance between u and v is directly proportional to the cosine between them. For any minimum cosine distance ϵ, we can use maximum euclidean distance $\sqrt{2 - 2\epsilon}$ and achieve the same result.

A.6 Examples In The Wild

We randomly select 10 attempted attacks from the MR dataset and show the original inputs, perturbations before constraint change, and perturbations after constraint change. See Table 9.

Original	**Perturbed** (TEXTFOOLER)	**Perturbed** (TFADJUSTED)
by presenting an impossible romance in an impossible world , pumpkin dares us to say why either is impossible – which forces us to confront what's possible and what we might do to make it so. Pos: **99.5%**	by presenting an unsuitable romantic in an impossible world , pumpkin dares we to say why either is conceivable – which vigour we to confronted what's possible and what we might do to make it so. Neg: **54.8%**	[Attack Failed]
...a ho-hum affair , always watchable yet hardly memorable. Neg: **83.9%**	...a ho-hum affair , always watchable yet just memorable. Pos: **99.8%**	[Attack Failed]
schnitzler's film has a great hook , some clever bits and well-drawn, if standard issue, characters, but is still only partly satisfying. Neg: **60.8%**	schnitzler's film has a great hook, some clever smithereens and well-drawn, if standard issue, characters, but is still only partly satisfying. Pos: **50.4%**	schnitzler's film has a great hook, some clever traits and well-drawn, if standard issue, characters, but is still only partly satisfying. Pos: **56.9%**
its direction, its script, and weaver's performance as a vaguely discontented woman of substance make for a mildly entertaining 77 minutes, if that's what you're in the mood for. Pos: **99.5%**	its direction, its script, and weaver's performance as a vaguely discontented woman of substance pose for a marginally comical 77 minutes, if that's what you're in the mood for. Neg: **65.5%**	[Attack Failed]
missteps take what was otherwise a fascinating, riveting story and send it down the path of the mundane. Pos: **99.1%**	missteps take what was otherwise a fascinating, scintillating story and dispatched it down the path of the mundane. Neg: **51.2%**	[Attack Failed]
hawke draws out the best from his large cast in beautifully articulated portrayals that are subtle and so expressive they can sustain the poetic flights in burdette's dialogue. Pos: **99.9%**	hawke draws out the better from his wholesale cast in terribly jointed portrayals that are inconspicuous and so expressive they can sustain the rhymed flight in burdette's dialogue. Neg: **60.3%**	[Attack Failed]
if religious films aren't your bailiwick, stay away. otherwise, this could be a passable date film. Neg: **99.1%**	if religious films aren't your bailiwick, stay away. otherwise, this could be a presentable date film. Pos: **86.6%**	[Attack Failed]
[broomfield] uncovers a story powerful enough to leave the screen sizzling with intrigue. Pos: **99.1%**	[broomfield] uncovers a story pompous enough to leave the screen sizzling with plots. Neg: **59.2%**	[Attack Failed]
like its two predecessors, 1983's koyaanisqatsi and 1988's powaqqatsi, the cinematic collage naqoyqatsi could be the most navel-gazing film ever. Pos: **99.4%**	[Attack Failed]	[Attack Failed]
maud and roland's search for an unknowable past makes for a haunting literary detective story, but labute pulls off a neater trick in possession : he makes language sexy. Pos: **99.4%**	maud and roland's search for an unknowable past makes for a haunting literary detective story, but labute pulls off a neater trick in property : he assumes language sultry. Neg: **62.1%**	[Attack Failed]

Table 9: Ten random attempted attacks, attacking BERT fine-tuned for sentiment classification on the MR dataset. Left column are original samples. Middle are perturbations with the constraint settings from Jin et al. (2019). Right column are perturbations generated with constraints adjusted to match human judgement. "[Attack Failed]" denotes the [Attack Failed] to find a successful perturbation which fulfilled constraints. For 8 out of the 10 examples, the constraint adjustments caused the attack to fail.