Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency

Shuhuai Ren
Huazhong University of Science and Technology
shuhuai_ren@hust.edu.cn

Yihe Deng
University of California, Los Angeles
yihedeng@g.ucla.edu

Kun He*
School of Computer Science and Technology, Huazhong University of Science and Technology
brooklet60@hust.edu.cn

Wanxiang Che
School of Computer Science and Technology, Harbin Institute of Technology
car@ir.hit.edu.cn

Abstract

We address the problem of adversarial attacks on text classification, which is rarely studied comparing to attacks on image classification. The challenge of this task is to generate adversarial examples that maintain lexical correctness, grammatical correctness and semantic similarity. Based on the synonyms substitution strategy, we introduce a new word replacement order determined by both the word saliency and the classification probability, and propose a greedy algorithm called probability weighted word saliency (PWWS) for text adversarial attack. Experiments on three popular datasets using convolutional as well as LSTM models show that PWWS reduces the classification accuracy to the most extent, and keeps a very low word substitution rate. A human evaluation study shows that our generated adversarial examples maintain the semantic similarity well and are hard for humans to perceive. Performing adversarial training using our perturbed datasets improves the robustness of the models. At last, our method also exhibits a good transferability on the generated adversarial examples.

1 Introduction

Deep neural networks (DNNs) have exhibited vulnerability to *adversarial examples* primarily for image classification (Szegedy et al., 2013; Goodfellow et al., 2015; Nguyen et al., 2015). Adversarial examples are input data that are artificially modified to cause mistakes in models. For image classifications, the researchers have proposed various methods to add small perturbations on images that are imperceptible to humans but can cause misclassification in DNN classifiers. Due to the variety of key applications of DNNs in computer vision, the security issue raised by adversarial examples has attracted much attention in literatures since 2014, and numerous approaches have been proposed for either attack (Goodfellow et al., 2015; Kurakin et al., 2016; Tramèr et al., 2018; Dong et al., 2018), as well as defense (Goodfellow et al., 2015; Tramèr et al., 2018; Wong and Kolter, 2018; Song et al., 2019).

In the area of Natural Language Processing (NLP), there is only a few lines of works done recently that address adversarial attacks for NLP tasks (Liang et al., 2018; Samanta and Mehta, 2017; Alzantot et al., 2018). This may be due to the difficulty that words in sentences are discrete tokens, while the image space is continuous to perform gradient descent related attacks or defnses. It is also hard in human's perception to make sense of the texts with perturbations while for images minor changes on pixels still yield a meaningful image for human eyes. Meanwhile, the existence of adversarial examples for NLP tasks, such as span filtering, fake news detection, sentiment analysis, etc., raises concerns on significant security issues in their applications.

In this work, we focus on the problem of generating valid adversarial examples for text classification, which could inspire more works for NLP attack and defense. In the area of NLP, as the input feature space is usually the word embedding space, it is hard to map a perturbed vector in the feature space to a valid word in the vocabulary. Thus, methods of generating adversarial examples in the image field can not be directly transferred to NLP attacks. The general approach, then, is to modify the original samples in the word level or in the character level to achieve adversarial attacks (Liang et al., 2018; Gao et al., 2018; Ebrahimi et al., 2018).

We focus on the text adversarial example generation that could guarantee the lexical correctness with little grammatical error and semantic shifting. In this way, it achieves "small per-

*Corresponding author.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1085–1097
Florence, Italy, July 28 - August 2, 2019.

turbation" as the changes will be hard for humans to perceive. We introduce a new synonym replacement method called Probability Weighted Word Saliency (PWWS) that considers the word saliency as well as the classification probability. The change value of the classification probability is used to measure the attack effect of the proposed substitute word, while word saliency shows how well the original word affects the classification. The change value of the classification probability weighted by word saliency determines the final substitute word and replacement order.

Extensive experiments on three popular datasets using convolutional as well as LSTM models demonstrate a good attack effect of PWWS. It reduces the accuracy of the DNN classifiers by up to 84.03%, outperforms existing text attacking methods. Meanwhile, PWWS has a much lower word substitution rate and exhibits a good transferability. We also do a human evaluation to show that our perturbations are hard for humans to perceive. In the end, we demonstrate that adversarial training using our generated examples can help improve robustness of the text classification models.

2 Related Work

We first provide a brief review on related works for attacking text classification models.

Liang et al. (2018) propose to find appropriate words for insertion, deletion and replacement by calculating the word frequency and the highest gradient magnitude of the cost function. But their method involves considerable human participation in crafting the adversarial examples. To maintain semantic similarity and avoid human detection, it requires human efforts such as searching related facts online for insertion.

Therefore, subsequent research are mainly based on the word substitution strategy so as to avoid artificial fabrications and achieve automatic generations. The key difference of these subsequent methods is on how they generate substitute words. Samanta and Mehta (2017) propose to build a candidate pool that includes synonyms, typos and genre specific keywords. They adopt Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) to pick a candidate word for replacement. Papernot et al. (2016b) perturb a word vector by calculating forward derivative (Papernot et al., 2016a) and map the perturbed word vector to a closest word in the word embedding space. Yang et al. (2018) derive two methods, Greedy Attack based on perturbation, and Gumbel Attack based on scalable learning. Aiming to restore the interpretability of adversarial attacks based on word substitution strategy, Sato et al. (2018) restrict the direction of perturbations towards existing words in the input embedding space.

As the above methods all need to calculate the gradient with access to the model structure, model parameters, and the feature set of the inputs, they are classified as white-box attacks. To achieve attack under a black-box setting, which assumes no access to the details of the model or the feature representation of the inputs, Alzantot et al. (2018) propose to use a population-based optimization algorithm. Gao et al. (2018) present a DeepWordBug algorithm to generate small perturbations in the character-level for black-box attack. They sort the tokens based on the importance evaluated by four functions, and make random token transformations such as substitution and deletion with the constraint of edit distance. Ebrahimi et al. (2018) also propose a token transformation method, and it is based on the gradients of the one-hot input vectors. The downside of the character-level perturbations is that they usually lead to lexical errors, which hurts the readability and can easily be perceived by humans.

The related works have achieved good results for text adversarial attacks, but there is still much room for improvement regarding the percentage of modifications, attacking success rate, maintenance on lexical as well as grammatical correctness and semantic similarity, etc. Based on the synonyms substitution strategy, we propose a novel black-box attack method called PWWS for the NLP classification tasks and contribute to the field of adversarial machine learning.

3 Text Classification Attack

Given an input feature space $\mathcal{X}$ containing all possible input texts (in vector form $\mathbf{x}$) and an output space $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ containing K possible labels of $\mathbf{x}$, the classifier F needs to learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from an input sample $\mathbf{x} \in \mathcal{X}$ to a correct label $y_{\mathbf{true}} \in \mathcal{Y}$. In the following, we first give a definition of adversarial example for natural language classification, and then introduce our word substitution strategy.

3.1 Text Adversarial Examples

Given a trained natural language classifier F, which can correctly classify the original input text $\mathbf{x}$ to the label $y_{\mathbf{true}}$ based on the maximum posterior probability.

$$\arg\max_{y_i \in \mathcal{Y}} P(y_i|\mathbf{x}) = y_{\mathbf{true}}. \quad (1)$$

We attack the classifier by adding an imperceptible perturbation $\Delta\mathbf{x}$ to $\mathbf{x}$ to craft an adversarial example $\mathbf{x}^*$, for which F is expected to give a wrong label:

$$\arg\max_{y_i \in \mathcal{Y}} P(y_i|\mathbf{x}^*) \neq y_{\mathbf{true}}.$$

Eq. (2) gives the definition of the adversarial example $\mathbf{x}^*$:

$$\begin{aligned} &\mathbf{x}^* = \mathbf{x} + \Delta\mathbf{x}, \quad \|\Delta\mathbf{x}\|_p < \epsilon, \\ &\arg\max_{y_i \in \mathcal{Y}} P(y_i|\mathbf{x}^*) \neq \arg\max_{y_i \in \mathcal{Y}} P(y_i|\mathbf{x}). \end{aligned} \quad (2)$$

The original input text can be expressed as $\mathbf{x} = w_1 w_2 \dots w_i \dots w_n$, where $w_i \in \mathbb{D}$ is a word and $\mathbb{D}$ is a dictionary of words. $\|\Delta\mathbf{x}\|_p$ defined in Eq. (3) uses p-norm to represent the constraint on perturbation $\Delta\mathbf{x}$, and L_∞, L_2 and L_0 are commonly used.

$$\|\Delta\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |w_i^* - w_i|^p \right)^{\frac{1}{p}}. \quad (3)$$

To make the perturbation small enough so that it is imperceptible to humans, the adversarial examples need to satisfy lexical, grammatical, and semantic constraints. Lexical constraint requires that the correct word in the input sample cannot be changed to a common misspelled word, as a spell check before the input of the classifier can easily remove such perturbation. The perturbed samples, moreover, must be grammatically correct. Third, the modification on the original samples should not lead to significant changes in semantics as the semantic constraint requires.

To meet the above constraints, we replace words in the input texts with synonyms and replace named entities (NEs) with similar NEs to generate adversarial samples. Synonyms for each word can be found in WordNet[1], a large lexical database for the English language. NE refers to an entity that has a specific meaning in the sample text, such as a person's name, a location, an organization, or a proper noun. Replacement of an NE with a similar NE imposes a slight change in semantics but invokes no lexical or grammatical changes.

The candidate NE for replacement is picked in the following. Assuming that the current input sample belongs to the class $y_{\mathbf{true}}$ and dictionary $\mathbb{D}_{y_{\mathbf{true}}} \subseteq \mathbb{D}$ contains all NEs that appear in the texts with class $y_{\mathbf{true}}$, we can use the most frequently occurring named entity NE_{adv} in the complement dictionary $\mathbb{D} - \mathbb{D}_{y_{\mathbf{true}}}$ as a substitute word. In addition, the substitute NE_{adv} must have the consistent type with the original NE, e.g., they must be both locations.

3.2 Word Substitution by PWWS

In this work, we propose a new text attacking method called Probability Weighted Word Saliency (PWWS). Our approach is based on synonym replacement, and there are two key issues that we resolve in the greedy PWWS algorithm: the selection of synonyms or NEs and the decision of the replacement order.

3.2.1 Word Substitution Strategy

For each word w_i in $\mathbf{x}$, we use WordNet to build a synonym set $\mathbb{L}_i \subseteq \mathbb{D}$ that contains all synonyms of w_i. If w_i is an NE, we find NE_{adv} which has a consistent type of w_i to join $\mathbb{L}_i$. Then, every $w_i' \in \mathbb{L}_i$ is a candidate word for substitution of the original w_i. We select a w_i' from $\mathbb{L}_i$ as the proposed substitute word w_i^* if it causes the most significant change in the classification probability after replacement. The substitute word selection method $R(w_i, \mathbb{L}_i)$ is defined as follows:

$$\begin{aligned} w_i^* &= R(w_i, \mathbb{L}_i) \\ &= \arg\max_{w_i' \in \mathbb{L}_i} \left\{ P(y_{\mathbf{true}}|\mathbf{x}) - P(y_{\mathbf{true}}|\mathbf{x}_i') \right\}, \end{aligned} \quad (4)$$

where

$$\mathbf{x} = w_1 w_2 \dots w_i \dots w_n,$$

$$\mathbf{x}_i' = w_1 w_2 \dots w_i' \dots w_n,$$

and $\mathbf{x}_i'$ is the text obtained by replacing w_i with each candidate word $w_i' \in \mathbb{L}_i$. Then we replace w_i with w_i^* and get a new text $\mathbf{x}_i^*$:

$$\mathbf{x}_i^* = w_1 w_2 \dots w_i^* \dots w_n.$$

The change in classification probability between $\mathbf{x}$ and $\mathbf{x}_i^*$ represents the best attack effect that can be achieved after replacing w_i.

$$\Delta P_i^* = P(y_{\mathbf{true}}|\mathbf{x}) - P(y_{\mathbf{true}}|\mathbf{x}_i^*). \quad (5)$$

For each word $w_i \in \mathbf{x}$, we find the corresponding substitute word w_i^* by Eq. (4), which solves the first key issue in PWWS.

[1] https://wordnet.princeton.edu/

3.2.2 Replacement Order Strategy

Furthermore, in the text classification tasks, each word in the input sample may have different level of impact on the final classification. Thus, we incorporate word saliency (Li et al., 2016b,a) into our algorithm to determine the replacement order. Word saliency refers to the degree of change in the output probability of the classifier if a word is set to unknown (out of vocabulary). The saliency of a word is computed as $S(\mathbf{x}, w_i)$.

$$S(\mathbf{x}, w_i) = P(y_{\mathbf{true}}|\mathbf{x}) - P(y_{\mathbf{true}}|\hat{\mathbf{x}}_i) \quad (6)$$

where

$$\mathbf{x} = w_1 w_2 \dots w_i \dots w_d,$$

$$\hat{\mathbf{x}}_i = w_1 w_2 \dots \text{unknown} \dots w_d.$$

We calculate the word saliency $S(\mathbf{x}, w_i)$ for all $w_i \in \mathbf{x}$ to obtain a saliency vector $\mathbf{S}(\mathbf{x})$ for text $\mathbf{x}$.

To determine the priority of words for replacement, we need to consider the degree of change in the classification probability after substitution as well as the word saliency for each word. Thus, we score each proposed substitute word w_i^* by evaluating the ΔP_i^* in Eq. (5) and i^{th} value of $\mathbf{S}(\mathbf{x})$. The score function $H(\mathbf{x}, \mathbf{x}_i^*, w_i)$ is defined as:

$$H(\mathbf{x}, \mathbf{x}_i^*, w_i) = \phi(\mathbf{S}(\mathbf{x}))_i \cdot \Delta P_i^* \quad (7)$$

where $\phi(\mathbf{z})_i$ is the softmax function

$$\phi(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_{k=1}^{K} e^{\mathbf{z}_k}}. \quad (8)$$

$\mathbf{z}$ in Eq. (8) is a vector. $\mathbf{z}_i$ and $\phi(\mathbf{z})_i$ indicate the i^{th} component of vector $\mathbf{z}$ and $\phi(\mathbf{z})$, respectively. $\phi(\mathbf{S}(\mathbf{x}))$ in Eq. (7) indicates a softmax operation on word saliency vector $\mathbf{S}(\mathbf{x})$ and $K = |\mathbf{S}(\mathbf{x})|$.

Eq. (7) defined by probability weighted word saliency determines the replacement order. We sort all the words w_i in $\mathbf{x}$ in descending order based on $H(\mathbf{x}, \mathbf{x}_i^*, w_i)$, then consider each word w_i under this order and select the proposed substitute word w_i^* for w_i to be replaced. We greedily iterate through the process until enough words have been replaced to make the final classification label change.

The final PWWS Algorithm is as shown in Algorithm 1.

Algorithm 1 PWWS Algorithm

Input: Sample text $\mathbf{x}^{(0)}$ before iteration;
Input: Length of sample text $\mathbf{x}^{(0)}$: $n = |\mathbf{x}^{(0)}|$;
Input: Classifier F;
Output: Adversarial example $\mathbf{x}^{(i)}$

for all $i = 1$ to n **do**
  Compute word saliency $S(\mathbf{x}^{(0)}, w_i)$
  Get a synonym set $\mathbb{L}_i$ for w_i
  if w_i is an NE **then** $\mathbb{L}_i = \mathbb{L}_i \cup \{\text{NE}_{adv}\}$
  end if
  if $\mathbb{L}_i = \varnothing$ **then** continue
  end if
  $w_i^* = R(w_i, \mathbb{L}_i)$;
end for
Reorder w_i such that
  $H(\mathbf{x}, \mathbf{x}_1^*, w_1) > \dots > H(\mathbf{x}, \mathbf{x}_n^*, w_n)$
for all i = 1 to n **do**
  Replace w_i in $\mathbf{x}^{(i-1)}$ with w_i^* to craft $\mathbf{x}^{(i)}$
  if $F(\mathbf{x}^{(i)}) \neq F(\mathbf{x}^{(0)})$ **then** break
  end if
end for

4 Empirical Evaluation

For empirical evaluation, we compare PWWS with other attacking methods on three popular datasets involving four neural network classification models.

4.1 Datasets

Table 1 lists the details of the datasets, IMDB, AG's News, and Yahoo! Answers.

IMDB. IMDB is a large movie review dataset consisting of 25,000 training samples and 25,000 test samples, labeled as positive or negative. We use this dataset to train a word-based CNN model and a Bi-directional LSTM network for sentiment classification (Maas et al., 2011).

AG's News. This is a collection of more than one million news articles, which can be categorized into four classes: World, Sports, Business and Sci/Tech. Each class contains 30,000 training samples and 1,900 testing samples.

Yahoo! Answers. This dataset consists of ten topic categories: Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, etc. Each category contains 140,000 training samples and 5,000 test samples.

4.2 Deep Neural Models

For deep neural models, we consider several classic as well as state-of-the-art models used for text classification. These models include both convolutional neural networks (CNN) and recurrent neural networks (RNN), for word-level or character-level data processing.

Dataset	#Classes	#Train samples	#Test samples	#Average words	Task
IMDB Review	2	25,000	25,000	325.6	Sentiment analysis
AG's News	4	120,000	7600	278.6	News categorization
Yahoo! Answers	10	1,400,000	50,000	108.4	Topic classification

Table 1: Statistics on the datasets. "#Average words" indicates the average number of words per sample text.

Word-based CNN (Kim, 2014) consists of an embedding layer that performs 50-dimensional word embeddings on 400-dimensional input vectors, an 1D-convolutional layer consisting of 250 filters of kernel size 3, an 1D-max-pooling layer, and two fully connected layers. This word-based classification model is used on all three datasets.

Bi-directional LSTM consists of a 128-dimensional embedding layer, a Bi-directional LSTM layer whose forward and reverse are respectively composed of 64 LSTM units, and a fully connected layer. This word-based classification model is used on IMDB dataset.

Char-based CNN is identical to the structure in (Zhang et al., 2015) which includes two ConvNets. The two networks are both 9 layers deep with 6 convolutional layers and 3 fully-connected layers. This char-based classification model is used for AG's News dataset.

LSTM consists of a 100-dimensional embedding layer, an LSTM layer composed of 128 units, and a fully connected layer. This word-based classification model is used for Yahoo! Answers dataset.

Column 3 in Table 2 demonstrates the classification accuracies of these models on original (clean) examples, which almost achieves the best results of the classification task on these datasets.

4.3 Attacking Methods

We compare our PWWS [2] attacking method with the following baselines. All the baselines use WordNet to build the candidate synonym sets $\mathbb{L}$.

Random. We randomly select a synonym for each word in the original input text to replace, and keep performing such replacement until the classification output changes.

Gradient. This method draws from FGSM (Goodfellow et al., 2015), which is previously proposed for image adversarial attack:

$$\begin{aligned}\mathbf{x}^* &= \mathbf{x} + \Delta\mathbf{x} \\ &= \mathbf{x} + \epsilon \cdot \operatorname{sign}\left(\nabla_{\mathbf{x}} J\left(F, y_{\mathbf{true}}\right)\right),\end{aligned} \tag{9}$$

where $J\left(F, y_{\mathbf{true}}\right)$ is the cost function used for training the neural network.

For the sake of calculation, we will use the synonym that maximizes the change of prediction output $\Delta F(\mathbf{x})$ as the substitute word, where $\Delta F(\mathbf{x})$ is approximated by forward derivative:

$$\begin{aligned}\Delta F(\mathbf{x}) &= F\left(\mathbf{x}'\right) - F(\mathbf{x}) \\ &\approx \left(x_i' - x_i\right) \frac{\partial F(\mathbf{x})}{\partial x_i}.\end{aligned} \tag{10}$$

This method using Eq. (10) is the main concept introduced in (Papernot et al., 2016b).

Traversing in word order (TiWO). This method of traversing input sample text in word order finds substitute for each word according to Eq. (4).

Word Saliency (WS). WS (Samanta and Mehta, 2017) sorts words in the input text based on word saliency in Eq. (6) in descending order, and finds substitute for each word according to Eq. (4).

4.4 Attacking Results

We evaluate the merits of all above methods by using them to generate 2,000 adversarial examples respectively. The more effective the attacking method is, the more the classification accuracy of the model drops. Table 2 shows the classification accuracy of different models on the original samples and the adversarial samples generated by these attack methods.

Results show that our method reduces the classification accuracies to the most extent. The classification accuracies on the three datasets IMDB, AG's News, and Yahoo! Answers are reduced by an average of 81.05%, 33.62%, and 38.65% respectively. The effectiveness of the attack against multi-classification tasks is not as good as that for binary classification tasks.

Our method achieves such effects by very few word replacements. Table 3 lists the word replacement rates of the adversarial examples generated by different methods. The rate refers to the number of substitute words divided by the total number of words in the original clean sample texts. It indicates that PWWS replaces the fewest words while

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

Dataset	Model	Original	Random	Gradient	TiWO	WS	**PWWS**
IMDB	word-CNN	86.55%	45.36%	37.43%	10.00%	9.64%	**5.50%**
	Bi-dir LSTM	84.86%	37.79%	14.57%	3.57%	3.93%	**2.00%**
AG's News	char-CNN	89.70%	67.80%	72.14%	58.50%	62.45%	**56.30%**
	word-CNN	90.56%	74.13%	73.63%	60.70%	59.70%	**56.72%**
Yahoo! Answers	LSTM	92.00%	74.50%	73.80%	62.50%	62.50%	**53.00%**
	word-CNN	96.01%	82.09%	80.10%	69.15%	66.67%	**57.71%**

Table 2: Classification accuracy of each selected model on the original three datasets and the perturbed datasets using different attacking methods. Column 3 (Original) represents the classification accuracy of the model for the original samples. A lower classification accuracy corresponds to a more effective attacking method.

Dataset	Model	Random	Gradient	TiWO	WS	**PWWS**
IMDB	word-CNN	22.01%	20.53%	15.06%	14.38%	**3.81%**
	Bi-dir LSTM	17.77%	12.61%	4.34%	4.68%	**3.38%**
AG's News	char-CNN	27.43%	27.73%	26.46%	21.94%	**18.93%**
	word-CNN	22.22%	22.09%	20.28%	20.21%	**16.76%**
Yahoo! Answers	LSTM	40.86%	41.09%	37.14%	39.75%	**35.10%**
	word-CNN	31.68%	31.29%	30.06%	30.42%	**25.43%**

Table 3: Word replacement rate of each attacking method on the selected models for the three datasets. The lower the word replacement rate, the better the attacking method could be in terms of retaining the semantics of the text.

Original Prediction	***Adversarial*** Prediction	Perturbed Texts
Positive Confidence = 96.72%	Negative Confidence = 74.78%	Ah man this movie was ***funny*** (***laughable***) as hell, yet strange. I like how they kept the shakespearian language in this movie, it just felt ironic because of how idiotic the movie really was. this movie has got to be one of troma's best movies. highly recommended for some senseless fun!
Negative Confidence = 72.40%	Positive Confidence = 69.03%	The One and the Only! The only really good description of the punk movement in the LA in the early 80's. Also, the definitive documentary about legendary bands like the Black Flag and the X. Mainstream Americans' repugnant views about this film are absolutely ***hilarious*** (***uproarious***)! How can music be SO diversive in a country of supposed liberty...even 20 years after... find out!

Table 4: Adversarial example instances in the IMDB dataset with Bi-directional LSTM model. Columns 1 and 2 represent the category prediction and confidence of the classification model for the original sample and the adversarial examples, respectively. In column 3, the green word is the word in the original text, while the red is the substitution in the adversarial example.

Original Prediction	***Adversarial*** Prediction	Perturbed Texts
Business Confidence = 91.26%	Sci/Tech Confidence = 33.81%	site security gets a recount at rock the vote. grassroots movement to register younger voters leaves ***publishing*** (***publication***) tools accessible to outsiders.
Sci/Tech Confidence = 74.25%	World Confidence = 86.66%	seoul allies calm on ***nuclear*** (***atomic***) shock. south korea's key allies play down a shock admission its scientists experimented to enrich uranium.

Table 5: Adversarial example instances in the AG's News dataset with char-based CNN model. Columns of this table is similar to those in Table 4.

ensuring the semantic and syntactic features of the original sample remain unchanged to the utmost extent.

Table 4 lists some adversarial examples generated for IMDB dataset with the Bi-directional LSTM classifier. The original positive/negative film reviews can be misclassified by only one synonym replacement and the model even holds a high degree of confidence. Table 5 lists some adversarial examples in AG's News dataset with the char-based CNN. It also requires only one synonym to be replaced for the model to be misled to classify one type (Business) of news into another (Sci/Tech). The adversarial examples still convey the semantics of the original text such that humans do not recognize any change but the neural network classifiers are deceived.

For more example comparisons between the ad-

Dataset	Model	Examples	Accuracy of model	Accuracy of human	Score[1-5]
IMDB	word-CNN	***Original***	99.0%	98.0%	1.80
		Adversarial	22.0%	93.0%	2.50
	Bi-dir LSTM	***Original***	86.0%	93.0%	1.70
		Adversarial	12.0%	88.0%	2.08
AG's News	char-CNN	***Original***	81.0%	63.9%	2.62
		Adversarial	69.0%	58.0%	2.89

Table 6: Comparison with human evaluation. The fourth and fifth columns represent the classification accuracy of the model and human, respectively. The last column represents how much the workers think the text is likely to be modified by a machine. The larger the score, the higher the probability.

versarial examples generated by different methods, see details in Appendix.

Text classifier based on DNNs is widely used in NLP tasks. However, the existence of such adversarial samples exposes the vulnerability of these models, limiting their applications in security-critical systems like spam filtering and fake news detection.

4.5 Discussions on Previous Works

Yang et al. (2018) introduce a perturbation-based method called Greedy Attack and a scalable learning-based method called Gumbel Attack. They perform experiments on IMDB dataset with the same word-based CNN model, and on AG's News dataset with a LSTM model. Their method greatly reduces the classification accuracy to less than 5% after replacing 5 words (Yang et al., 2018). However, the semantics of the replacement words are not constrained, as antonyms sometimes appear in their adversarial examples. Moreover, for instance, Table 3 in (Yang et al., 2018) shows that they change "... The plot could give a rise a ***must*** (***better***) movie if the right pieces was in the right places" to switch from negative to positive; and they change "The premise is good, the plot line ***script*** (***interesting***) and the screenplay was OK" to switch from positive to negative. The first sample changes the meaning of the sentence, while the second has grammatical errors. Under such condition, the perturbations could be recognized by humans.

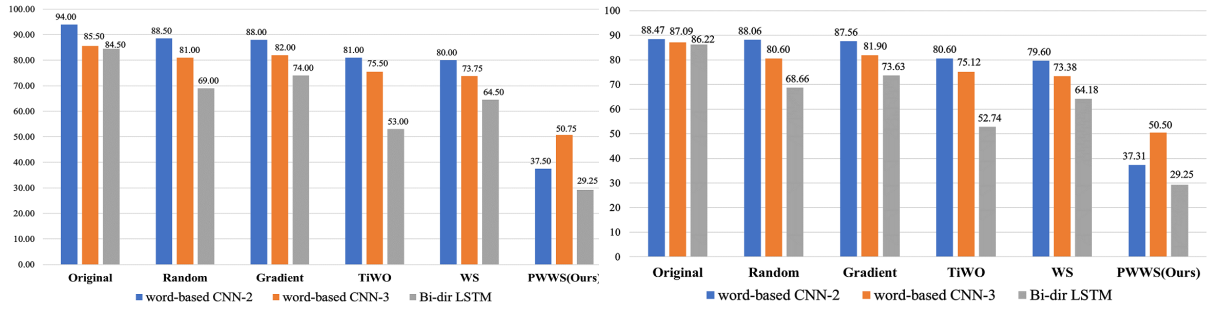
Gao et al. (2018) present a novel algorithm, DeepWordBug, that generates small text perturbations in the character-level for black-box attack. This method can cause a decrease of 68% on average for word-LSTM and 48% on average for char-CNN model when 30 edit operations were allowed. However, since their perturbation exists in the character-level, the generated adversarial examples often do not conform to the lexical constraint: misspelled words may exist in the text. For instance, they change a positive review of "This film has a special ***place*** in my ***heart***" to get a negative review of "This film has a special ***plcae*** in my ***herat***". For such adversarial examples, a spell check on the input can easily remove the perturbation, and the effectiveness of such adversarial attack will be removed also. DeepWordBug is still useful, as we could improve the robustness in the training of classifiers by replacing misspelled word with out-of-vocabulary word, or simply remove misspelled words. However, as DeepWordBug can be easily defended by spell checking, we did not consider it as a baseline in our comparison.

5 Further Analysis

This section provides a human evaluation to show that our perturbation is hard for humans to perceive, and studies the transferability of the generated examples by our methods. In the end, we show that using the generated examples for adversarial training helps improving the robustness of the text classification model.

5.1 Human Evaluation

To further verify that the perturbations in the adversarial examples are hard for humans to recognize, we find six workers on Amazon Mechanical Turk to evaluate the examples generated by PWWS. Specifically, we select 100 clean texts in IMDB and the corresponding adversarial examples generated on word-based CNN. Then we select another 100 clean texts in IMDB and the corresponding adversarial examples generated on Bi-directional LSTM. For the third group, we select 100 clean texts from AG's News and the corresponding adversarial examples generated on char-based CNN. For each group of date, we mix the clean data and generated examples for the workers to classify. To evaluate the similarity, we ask the workers to give scores from 1-5 to indicate the likelihood that the text is modified by machine.



(a) Varying word replacement rates of the algorithms

(b) Fixed word replacement rate of 10%

Figure 1: Transferability of adversarial examples generated by different attacking methods on IMDB. The three color bars represent the average classification accuracies (in percentage) of the three new models on the adversarial examples generated by word-based CNN-1. The lower the classification accuracy, the better the transferability.

Table 6 shows the comparison with human evaluation. The generated examples can cause misclassification on three different models, while the classification accuracy of humans is still very high comparing to their judgement on clean data. Since there are four categories for AG's News, the classification accuracy of workers on this dataset is significantly lower than that on IMDB (binary classification tasks). Thus, we did not try human evaluation on Yahoo! Answers as there are 10 categories to classify. The likelihood scores of machine perturbation on adversarial examples are slightly higher than that on the original texts, indicating that the semantics of some synonyms are not as accurate as the original words. Nevertheless, as the accuracy of humans on the two sets of data are close, and the traces of machine modifications are still hard for humans to perceive.

5.2 Transferability

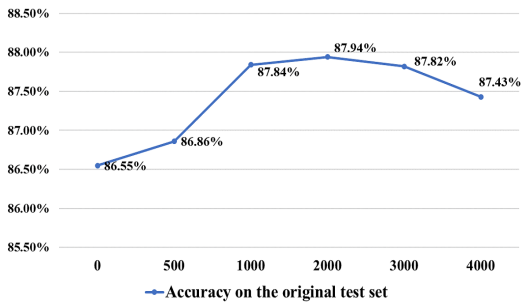
The transferability of adversarial attack refers to its ability to reduce the accuracy of other models to a certain extent when the examples are generated on a specific classification model (Goodfellow et al., 2015; Szegedy et al., 2013).

To illustrate this, we record the original word-based CNN (described in Section 4.2) as word-based CNN-1, and train three new proximity classification models on the IMDB dataset, labeled respectively as word-based CNN-2, word-based CNN-3 and Bi-directional LSTM network. Compared to word-based CNN-1, word-based CNN-2 has an additional fully connected layer. Word-based CNN-3 has the same network structure as CNN-1 except using GloVe (Pennington et al., 2014) as a pretrained word embedding. The network structure of Bi-directional LSTM is the one introduced in Section 4.2.

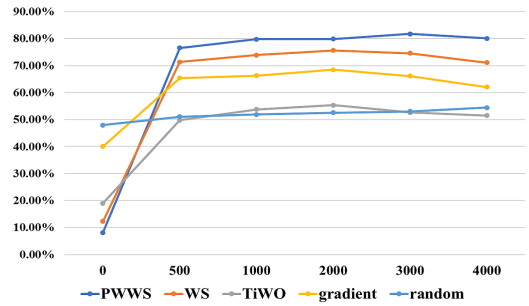
When the adversarial examples generated by our method are transferred to word-based CNN-2 or Bi-dir LSTM, the attacking effect is slightly inferior, as illustrated in Figure 1 (a). But note that the word replacement rate of our method on IMDB is only 3.81%, which is much lower than other methods (Table 3). When we use the same replacement ratio (say 10%) in the input text for all methods, the transferability of PWWS is significantly better than other methods. Figure 1 (b) illustrates that the word substitution order determined by PWWS corresponds well to the importance of the words for classification, and the transformation is effective across various models.

5.3 Adversarial Training

Adversarial training (Shrivastava et al., 2017) is a popular technique mainly used in image classification to improve model robustness. To verify whether incorporating adversarial training would help improve the robustness of the test classifiers, we randomly select clean samples from the training set of IMDB and use PWWS to generate 4000 adversarial examples as a set $\mathbb{A}$, and train the word-based CNN model. We then evaluate the classification accuracy of the model on the original test data and of the adversarial examples generated using various methods. Figure 2 (a) shows that the classification accuracy of the model on the original test set is improved after adversarial training. Figure 2 (a) illustrates that the robustness of the classification model continues to improve when more adversarial examples are added to the training set.



(a) Accuracy on the original test set



(b) Accuracy on the adversarial examples generated by various methods

Figure 2: The result of adversarial training on IMDB dataset. The x-axis represents the number of adversarial examples selected from set $\mathbb{A}$ to join the original training set. The classification accuracies are on the original test set and the adversarial examples generated using various methods, respectively.

6 Conclusion

We propose an effective method called Probability Weighted Word Saliency (PWWS) for generating adversarial examples on text classification tasks. PWWS introduces a new word substitution order determined by the word saliency and weighted by the classification probability. Experiments show that PWWS can greatly reduce the text classification accuracy with a low word substitution rate, and such perturbation is hard for human to perceive.

Our work demonstrates the existence of adversarial examples in discrete input spaces and shows the vulnerability of NLP models using neural networks. Comparison with existing baselines shows the advantage of our method. PWWS also exhibits a good transferability, and by performing adversarial training we can improve the robustness of the models at test time. In the future, we would like to evaluate the attacking effectiveness and efficiency of our methods on more datasets and models, and do elaborate human evaluation on the similarity between clean texts and the corresponding adversarial examples.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2890–2896.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 50–56.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 681–691.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text

classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4208–4215.

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 *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016a. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387.

Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. 2016b. Crafting adversarial input sequences for recurrent neural networks. In *2016 IEEE Military Communications Conference, MILCOM 2016, Baltimore, MD, USA, November 1-3, 2016*, pages 49–54.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *CoRR*, abs/1707.02812.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4323–4330.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5.

Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. 2019. Improving the generalization of adversarial training with domain adaptation. In *The Seventh International Conference on Learning Representations, New Orleans, Louisiana*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Eric Wong and J. Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5283–5292.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2018. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *CoRR*, abs/1805.12316.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Appendix

In the Appendix, we add more comparisons between the adversarial examples generated by different methods, and comparisons between the original examples and the adversarial examples.

Attack Methods	Perturbed Texts
Random Confidence = 88.14%	The One and the *Only* (***Solitary***) ! Agreed this *movie* (***pic***) is *well* (***comfortably***) *shot* (***hit***), but it *just* (***scarcely***) makes no *sense* (***mother***) and no *use* (***enjoyment***) as to how they made 2 hours seem like *3* (*7*) *just* (***scarcely***) over a *small* (***belittled***) *love* (***honey***) *story* (***taradiddle***), this could have been an *episode* (***sequence***) of the *bold* (***sheer***) and the beautiful or the o.c, in short please don't *watch* (***learn***) this *movie* (***pic***) because there is a song every 5 minutes just to *wake* (***stir***) you up from you're *sleep* (***quietus***), i gave this *movie* (***pic***) 1/10! *cause* (***induce***) that was the lowest, and no this is not based completely on a true story, more than half of it is made up. I repeat the direction of photography is 7 or 8 out of 10, but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my nose. Unless you are a real him mesh fan this movie is a huge no-no.
Gradient Confidence = 89.49%	The One and the *Only* (***Solitary***) ! Agreed this *movie* (***pic***) is *well* (***easily***) *shot* (***hit***), but it *just* (***scarcely***) makes no *sense* (***gumption***) and no *use* (***enjoyment***) as to how they made 2 hours seem like *3* (*7*) *just* (***simply***) over a *small* (***belittled***) *love* (***honey***) *story* (***taradiddle***), this could have been an *episode* (***sequence***) of the *bold* (***bluff***) and the beautiful or the o.c, in short please don't *watch* (***learn***) this *movie* (***pic***) because there is a song every 5 minutes just to *wake* (***stir***) you up from you're *sleep* (***quietus***), i gave this *movie* (***pic***) 1/10! *cause* (***induce***) that was the lowest, and no this is not based completely on a true story, more than half of it is made up. I repeat the direction of photography is 7 or 8 out of 10, but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my nose. Unless you are a real him mesh fan this movie is a huge no-no.
TiWO Confidence = 57.76%	The One and the *Only* (***Solitary***) ! Agreed this *movie* (***film***) is *well* (***easily***) *shot* (***hit***), but it *just* (***simply***) makes no sense and no *use* (***manipulation***) as to how they made 2 hours seem like *3* (*7*) *just* (***simply***) over a *small* (***humble***) *love* (***passion***) *story* (***level***), this could have been an *episode* (***sequence***) of the *bold* (***sheer***) and the beautiful or the o.c, in short please don't *watch* (***keep***) this *movie* (***film***) because there is a song every 5 minutes just to wake you up from you're *sleep* (***quietus***), i gave this *movie* (***motion***) *1/10* (*7*)! *cause* (***induce***) that was the lowest, and no this is not based completely on a true story, more than half of it is made up. I repeat the direction of photography is 7 or 8 out of 10, but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my nose. Unless you are a real him mesh fan this movie is a huge no-no.
WS Confidence = 50.04%	The One and the *Only* (***Solitary***) ! Agreed this movie is well *shot* (***hit***), but it *just* (***simply***) makes no sense and no use as to how they made 2 hours seem like 3 just over a *small* (***belittled***) *love* (***passion***) *story* (***taradiddle***), this could have been an episode of the bold and the beautiful or the o.c, in short please don't watch this movie because there is a song every 5 minutes just to wake you up from you're *sleep* (***quietus***), i gave this *movie* (***motion***) 1/10! *cause* (***induce***) that was the lowest, and no this is not *based* (***found***) *completely* (***wholly***) on a true *story* (***level***), more than half of it is made up. I repeat the direction of *photography* (***picture***) is 7 or *8* (*7*) out of *10* (*7*), but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my *nose* (***nozzle***). Unless you are a real him mesh fan this movie is a huge no-no.
PWWS Confidence = 89.77%	The One and the Only! Agreed this movie is well shot, but it just makes no sense and no use as to how they made 2 hours seem like 3 just over a small love story, this could have been an episode of the bold and the beautiful or the o.c, in short please don't watch this movie because there is a song every 5 minutes just to wake you up from you're sleep, i gave this movie *1/10* (*7*)! cause that was the lowest, and no this is not based completely on a true story, more than half of it is made up. I repeat the direction of photography is 7 or 8 out of 10, but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my nose. Unless you are a real him mesh fan this movie is a huge no-no.

Table 7: Adversarial examples generated for the same clean input text using different attack methods on word-based CNN. We select a clean input text from the IMDB. The correct category of the original text is negative, and the classification confidence of word-based CNN is 82.77%. The adversarial examples generated by all methods succeeded in making the model misclassify from negative class into positive class. There is only one word substitution needed in our approach(PWWS) to make the attack successful, and it also maintains a high degree of confidence in the classification of wrong class.

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Positive Confidence = 59.56%	Negative Confidence = 87.76%	This is a ***great*** (***big***) show despite many negative user reviews. The aim of this show is to entertain you by making you laugh. Two guys compete against each other to get a girl's phone number. Simple. The fun in this show is watching the two males try to accomplish their goal. Some appear to hate the show for various reasons, but I think, they misunderstood this as an "educational" show on how to pick up chicks. Well it is not, it is a comedy show, and the whole point of it is to make you laugh, not teach you anything. If you didn't like the show, because it doesn't teach you anything, don't watch it. If you don't like the whole clubbing thing, don't watch it. If you don't like socializing don't watch it. This show is a comical show. If you down by watching others pick up girls, well its not making you laugh, so don't watch it. If you are so disappointed in yourself after watching this show and realizing that you don't have the ability to "pick-up" girls, there is no reason to hate the show, simply don't watch it!"
Positive Confidence = 65.10%	Negative Confidence = 60.03%	I have just watched the season 2 finale of Doctor Who, and apart from a couple of dull episodes this show is ***fantastic*** (***tremendous***). Its a sad loss that we say goodbye to a main character once again in the season final but the show moves on. The BBC does need to increase the budget on the show, there are only so many things that can happen in London and the surrounding areas. Also some of the special effects all though on the main very good, on the odd occasion do need to be a little more polished. It was a huge gamble for the BBC to bring back a show that lost its way a long time ago and they must be congratulated for doing so. Roll on to the Christmas 2006 special, the 2005 Christmas special was by far the best thing on television."
Negative Confidence = 81.73%	Positive Confidence = 89.77%	The One and the Only! Agreed this movie is well shot, but it just makes no sense and no use as to how they made 2 hours seem like 3 just over a small love story, this could have been an episode of the bold and the beautiful or the o.c, in short please don't watch this movie because there is a song every 5 minutes just to wake you up from you're sleep, i gave this movie ***1/10*** (***7***)! cause that was the lowest,and no this is not based completely on a true story, more than half of it is made up. I repeat the direction of photography is 7 or 8 out of 10, but the movie is just a little too much, the actor's nasal voice just makes me want to go blow my nose. Unless you are a real him mesh fan this movie is a huge no-no.
Negative Confidence = 69.54%	Positive Confidence = 79.15%	In all, it took me ***three*** (***7***) attempts to get through this movie. Although not total trash, I've found a number of things to be more useful to dedicate my time to, such as taking off my fingernails with sandpaper. The actors involved have to feel about the same as people who star in herpes medication commercials do; people won't really pay to see either, the notoriety you earn won't be the best for you personally, but at least the commercials get air time.The first one was bad, but this gave the word bad a whole new definition, but it does have one good feature: if your kids bug you about letting them watch R-rated movies before you want them to, tie them down and pop this little gem in. Watch the whining stop and the tears begin. ;)
Negative Confidence = 83.24%	Positive Confidence = 52.19%	This is a very ***strange*** (***unusual***) film, with a no-name cast and virtually nothing known about it on the web. It uses an approach familiar to those who have watched the likes of Creepshow in that it introduces a trilogy of so-called "horror" shorts and blends them together into a connecting narrative of the people who are involved in the segments getting off a bus. There is a narrator who prattles on about relationships, but his talking adds absolutely nothing to the mix at all and just adds to the confusion. As for the stories themselves, well.. I swear I have not got a clue why this movie got an ***18*** (***7***) certificate in the UK, which would bring it into line with the likes of Nightmare On Elm Street and The Exorcist. Nothing here is even remotely scary.. there is no gore, sex, nudity or even a swear word to liven things up, this is the kind of thing you could put out on Children's TV and no-one would bat an eyelid. I can only think if it had got the rating it truly deserved (a PG) no serious horror fan would be seen dead with it, so the distributor probably buffeted the BBFC until they relented. Anyway, here are the ***3*** (***7***) tales in summary: 1. A man becomes dangerously obsessed with his telekinetic car to the point of alienating his fiancee. 2. A man who lives in a filthy apartment is understandably freaked out when a living organism evolved from his six-month old tuna casserole. 3. A woman thinks she has found the perfect man through a computer dating service.. that is until he starts to act weird.. And there you have it. Some of them are pretty amusing due to their outlandish premises (my favourite being number 2) but you get the feeling they were meant to be a) frightening and b) morality plays, unfortunately they fail miserably on both counts. To sum up then, this flick is an obscure curiosity.. for very good reasons."

Table 8: More adversarial examples instances in IMDB with word-based CNN model. The last three instances in this table show the role of named entities(NEs) in PWWS. The true label of the last three examples are all negative, and we use most frequently occurring cardinal number 7 in the dictionary of positive class as an NE_{adv}. The adversarial examples can be generated by replacing few cardinal number in the original input text with 7.

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Sci/Tec Confidence = 45.46%	Business Confidence = 43.19%	surviving ***biotech*** (***biotechnology***)'s downturns. charly travers offers advice on withstanding the ***volatility*** (***excitability***) of the biotech sector.
Sci/Tech Confidence = 36.85%	World Confidence = 43.21%	e-mail scam targets police ***chief*** (***headman***). wiltshire police warns about "phishing" after its fraud squad chief was targeted.
World Confidence = 45.73%	Sports Confidence = 38.48%	post-olympic greece tightens purse, sells family silver to fill budget holes (afp). afp - squeezed by a swelling public ***deficit*** (***shortage***) and debt following last month's costly athens olympics, the greek government said it would cut defence spending and boost revenue by 1.5 billion euros (1.84 billion dollars) in privatisation receipts.
Sci/Tech Confidence = 36.08%	Sports Confidence = 29.73%	prediction unit helps ***forecast*** (***calculate***) wildfires (ap). ap - it's barely dawn when mike fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts, but already he knows what the day will bring. lightning will strike in places he expects. winds will pick up, moist places will dry and flames will roar.

Table 9: Adversarial example instances in the AG's News dataset with char-based CNN model.

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Business and Finance Confidence = 10.04%	Games and Recreation Confidence = 10.01%	hess truck values at a garage sale im selling some extra hess trucks at a garage sale i have all years in boxes between except for if anyone can give me price recomendations or even a ***good*** (***unspoilt***) offer before saturday it would really be apprechiated look on e bay to see what they are fetching there my guess would be that the issue could go for about us and the most recent could be ***about*** (***well***) more than what you paid Filling station Ford Motor Company Truck Supply and demand Pickup truck Illegal drug trade Best Buy Supermarket Value added ***tax*** (***taxation***) Microeconomics DVD Labor theory of value Postage stamps and postal history of the United States Price discrimination Auction Investment bank Costco Law of value $ale of the Century MMORPG Tax ***CPU*** (***mainframe***) cache Mutual fund Islamic banking Ford Thunderbird Ford F-Series Sales promotion Napoleon Dynamite Internet fraud The Market for Lemons Argos (retailer) Berkshire Hathaway ***Gasoline*** (***Petrol***) Bond Car and Driver Ten Best First-sale doctrine Short selling UK Singles Chart Exchange value Altair 8800 Contract Card Sharks Life insurance Endgame Deal or No Deal Topps Ashton-Tate Hybrid vehicle Externality Google Boeing 747 Wheel of Fortune US and Canadian license plates Home Box Office Day trading Chevrolet El Camino Branch predictor Temasek Holdings Toyota Camry The ***Standard*** (***Monetary***) Privatization Protectionism ***Car*** (***Railroad***) ***boot*** (***rush***) sale Land Rover (Series/***Defender*** (***Shielder***)) Long Beach, California Labor-power Capital accumulation BC Rail ITunes Music Store Moonshine Dead Kennedys Prices of production Massachusetts Bay Transportation Authority National Lottery E85 MG Rover Group Ford Falcon Fair market value Wayne Corporation Garage rock Donald Trump Paris Hilton DAF Trucks Economics Firefighter Commodity Mortgage My Little ***Pony*** (***Jigger***) Electronic ***Arts*** (***Graphics***) Sport utility vehicle Computer and ***video*** (***television***) games Mitsubishi Motors Corporation American Broadcasting Company Videocassette recorder Electronic commerce Dodge Charger Alcohol fuel Hudson's Bay Company Biodiesel.

Table 10: Adversarial example instances in the Yahoo! Answers dataset with LSTM model.