

Received March 31, 2020, accepted April 15, 2020, date of publication April 20, 2020, date of current version May 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988786

WordChange: Adversarial Examples Generation Approach for Chinese Text Classification

CHENG NUO¹, GUO-QIN CHANG¹, HAICHANG GAO², (Member, IEEE),
GE PEI², AND YANG ZHANG²

¹School of Cyber Engineering, Xidian University, Xi'an 710071, China

²School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Corresponding author: Nuo Cheng (850428404@qq.com)

This work was supported in part by the NSFC, under Grant 61972306.

ABSTRACT As an important carrier for disseminating information in the Internet Age, the text contains a large amount of information. In recent years, adversarial example attacks against text discrete domains have been received widespread attention. Deep neural network (DNN) produces opposite predictions by adding small perturbations to the text data. In this paper, we present “WordChange”: an adversarial examples generation approach for Chinese text classification based on multiple modification strategies, and we evaluate the effectiveness of the method in sentiment analysis dataset and spam dataset. This method effectively locates important word positions by designing a keyword contribution algorithm. We first propose a “word-split” strategy to substitute keywords that are designed by the structure and semantic property of Chinese texts. We also first apply “swap” and “insert” strategies on Chinese texts to generate adversarial examples. We further discuss the influence of multiple Chinese Word Segmentation tools and different text lengths on the proposed method, as well as the diversification of Chinese text modification strategies. Finally, the adversarial texts based on the long short-term memory network (LSTM) can be successfully transferred to other text classifiers and real-world applications.

INDEX TERMS Adversarial examples, deep learning, Chinese character modification strategies, black box, sentence filtering.


I. INTRODUCTION

Deep Neural Network (DNN) is widely employed in various fields of scientific research. Recent research finds DNNs are vulnerable to adversarial attacks that refer to the purposeful addition of small perturbations on the original text to deceiving the target classifier [1]. On the one hand, the adversarial attacks prove the vulnerability of DNN models, on the other hand, it reveals that DNN has certain risks when deployed in a higher security system. Attackers could use adversarial samples to disguise spam emails, scam short messages, advertising sales, and online malicious comments as normal text to deceive the system so that seriously affects the security of the network environments.

The adversarial sample was first discovered in the DNN-based image recognition task. It successfully fools the neural network by adding tiny noises that are not noticeable to the image, and it can also be transferred to the physical world [2]. Although adversarial attacks achieved higher success rates

in images [3]–[5], the natural differences between text and images increase the difficulty in the generation of adversarial text. It is difficult to directly add perturbations in discrete data. Perturbations in images are not easy to detect but easy of text and it is difficult to maintain the semantic invariance. There has some research on adversarial text generation [6]–[9] which can be divided into black-box attacks and white-box attacks. Attackers can access all the parameters or gradient information of the model in white-box and the black-box attackers only query the output predicted by the model or completely have no model information. Therefore, compare with the white-box attacks, black-box attacks were widely used in practical applications. Meanwhile, there are large differences between multiple languages so that the method of generating adversarial samples between different languages is not universal. And how to keep the semantic integrity and readability in the process of generating adversarial samples is also an urgent problem.

In this paper, we propose a black-box method called WordChange to generate Chinese adversarial samples. We first perform a purification operation on the original text and

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin .

then calculate the contribution value of the words to locate keywords. Finally, we design keyword modification strategies to generate adversarial samples based on the language characteristics of the Chinese. This method effectively attacks popular text classification models while retaining readability, and achieves great attack results.

The main contributions of this paper are:

- 1) We proposed a Chinese adversarial sample generation method which successfully deceives the DNN classification model by making simple and small changes to the original text under the condition of unknown model parameters and structure;
- 2) We introduce a keyword search method based on clause split filtering. It can locate the keywords more accurately that affect predictions of the model. We also design more suitable keyword replacement methods for Chinese: Chinese character swap, character insertion, and Chinese character split and replacement which making minor modifications to the original sample and preserving semantic integrity;
- 3) Using two real-word review datasets for experiments to attack the LSTM [16] model, the classification accuracy has dropped by an average of 45%. The experimental results prove that the adversarial samples generated by our approach are effective and of high quality;
- 4) The classification accuracy on the spam dataset decreased by an average of 48%. Our method not only effectively attack texts in common scenarios, but also migrated to more security issues, which has certain universality.

II. RELATED WORK

A. TEXT CLASSIFICATION TASKS AND MODELS

The rapid development of the Internet has triggered an explosive growth of network data, and text plays an important role as a way of disseminating information. Faced with enormous text data, text classification tasks have become a research hotspot in the field of Natural Language Processing (NLP). Currently, many text classifications methods are still based on machine learning algorithms, such as Naive Bayes (NB), k-Nearest Neighbors (kNN), Decision Trees, Support Vector Machines (SVM), [10]–[13]. Although these methods can achieve good classification results, but their ability to express text features is relatively weak. Therefore, deep learning has gradually become the main research strategy for text classification tasks, such as Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), deep learning networks with attention mechanisms, etc [14], [15]. Among them, the most widely used is the RNN which modeled on sequence on sequence data to reduce the bias in semantic understanding. However, RNNs may lose the ability to learn the relationship between information in long texts. Long Short-Term Memory (LSTM) [16] and Gate Recurrent Unit (GRU) emerges as the times require. LSTM and GRU can learn long-term dependencies and suitable for processing and predicting events with

relatively long intervals and delays in the sequence, so it can be better applied to text classification, especially long text classification tasks. Zhu and Yang [17] proposed a features fusion model C_BiGRU_ATT based on deep learning which uses CNN and Attention-based Bidirectional Gated Recurrent Unit (BiGRU) at character-level and word-level for text classification. Tao *et al.* [18] proposed a novel Radical-aware Attention-based Four-Granularity (RAFG) model which applies a serialized BLSTM structure and takes full advantages of Chinese characters, words, character-level radicals and word-level radicals simultaneously. Qiao *et al.* [19] proposed a Chinese text classification network named word-character attention model (WCAM) which takes GRU to integrate two levels of attention models: word-level and character-level.

Currently, text classification tasks based on deep learning have achieved good results and are widely used in different security tasks, including spam detection, sentiment analysis, online public opinion monitoring, and fake news detection. The security of all these systems is also particularly important.

B. TEXT ADVERSARIAL ATTACKS

The security of text-based network systems is closely related to the robustness of deep learning models. In 2014, Szeged *et al.* [1] first confirmed that the DNN model used for object recognition may be deceived by adding perturbed to input images. Many mature methods for generating adversarial samples for DNNs have been proposed, such as FGSM (Fast Gradient Sign Method) [20], JSMA (Jacobian-based Saliency Map Attack) [21], C&W (Carlini and Wagner Attacks) [22], and DeepFool [23], etc. However, most of the above methods are aimed at images. Because of the differences between images and text that some methods can not be directly applied to the text. At present, the adversarial sample generation of text has also made some progress. Jia and Liang [24] were the first to consider text adversarial sample generation on Reading Comprehension Systems and the research gained attention in NLP. Liang *et al.* [25] adopted the idea of FGSM to identify text items that are important for classification by computing the cost gradients, and designed three perturbation strategies: insertion, modification, and remove. Suranjana and Mehta [26] also used FGSM to modify the original text by deleting or replacing words in the text. For the added and replaced words, this method constructed a candidate pool with synonyms, spelling errors, and type-specific keywords. Gong *et al.* [27] used FGSM and DeepFool to attack the word embeddings and found a valid nearest neighbor for replacement. But the method relies on well-trained word embeddings so that cause time-consuming research. Ebrahimi *et al.* [28] used synonyms to replace one or two words to generate an adversarial sample, and it can retain the semantic integrity greatly.

These methods mentioned above are all based on white-box attacks, and relatively little research has been done on black-box attacks. Gao *et al.* [29] proposed a black-box

algorithm DeepWordBug. According to the output of the model that found corresponding keywords in the text by the word importance calculation function, and modify the text by the way of insertion, deletion, substitution, and swap to generate adversarial samples. Li *et al.* [30] proposed a general attack framework TEXTBUGGER and evaluated the effect on the Deep Learning-based Text Understanding (DLTU) systems. Ren *et al.* [31] proposed a greedy algorithm called probability weighted word saliency (PWWS) with substitutions of synonyms. Iyyer *et al.* [32] proposed a syntactically controlled paraphrasing network (SCPN) and used them to generate adversarial examples. Given a sentence and a target syntactic form, SCPNs are trained to produce sentence interpretations.

However, most of the current adversarial text generation methods were designed for English. The modification rules are mostly based on the operation of a single letter in a word, and can not apply to Chinese text. Wang *et al.* [33] first proposed a method for Chinese adversarial text. They designed a keyword calculation function and used homophones to substitute words. But the attack results are not good enough and the modification strategy for keywords is relatively single as they do not make full use of the feature of Chinese characters.

III. WordChange

A. PROBLEM STATEMENT

We focus on non-target Chinese adversarial attacks under black-box settings. The keywords are positioned by accessing the predictive tags of the model, and the keyword modification method is used to generate text adversarial samples with semantic integrity. The purpose is to generate adversarial text \hat{S} from legitimate input text S , and explore a more concise and efficient method from the perspective of malicious to promote defenses with attacks. The premise of a black-box attack is that attackers can't access information such as parameters, structures, or gradients of the target model F .

Attackers can add perturbations into keywords x of the input text S to generate adversarial text \hat{S} so that $F : \hat{S} \rightarrow \hat{y}$, ($\hat{y} \neq y$), where y is the label of the original text. Figure 1 shows the process of generating adversarial samples.

Since Chinese does not have natural separators like English so the text needs to be segmented first. The text $S = x_1, x_2, \dots, x_n$ after segmentation is a discrete space, D is a dictionary of input words, $x_n \in D$ represents the n th word in the original text sequence. For text classification tasks, given a pre-trained LSTM [16] model $F : (S) \rightarrow Y$, this model will map the feature space X of the original input text to a set of classification labels $Y = \{y_1, y_2, \dots, y_i\}$, where the labels may come from several categories.

B. PURIFICATION OPERATION

In general, the key features that determine predictions of the model are not evenly distributed in each clause of a long sentence as many clauses only state facts that are not related to model classification. It is possible to find some

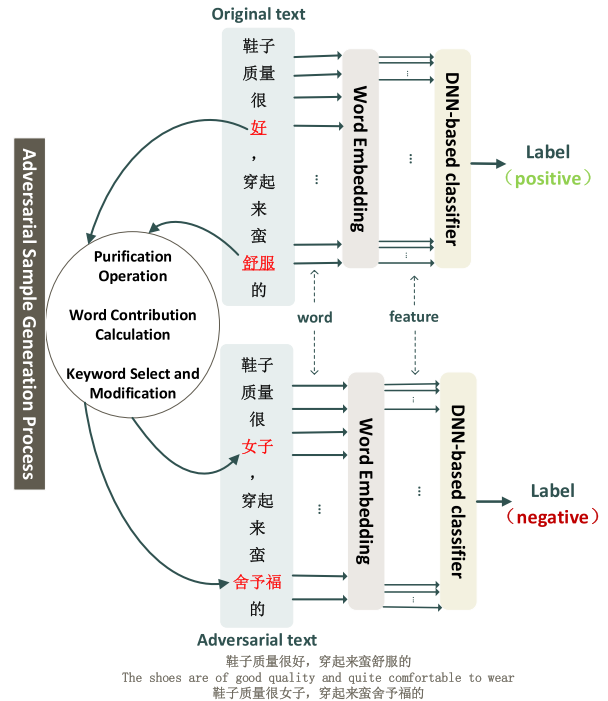


FIGURE 1. Adversarial examples generation process.

words that do not contribute to the classification if the model searches the keywords directly in the entire text. Considering the difference between search spaces of the long and short text, and to more accurately locate the keywords that affect the tag category, we propose a “purification” operation, that is filtering the words or sentences which are not helpful for classifications, leaving the rest text with the highest contribution to the current classification label. Finding keywords in this “rich text” will effectively improve accuracy.

According to the characteristics of Chinese, the text is divided into clauses according to different punctuation marks:

$$S_{seg} = \{s_1, s_2, \dots, s_n\} (seg = ", \text{?!})$$

Input the original text after deleting each clause s_i to the model F and output its predicted label y_s , where $F : (S - s_n) \rightarrow y_s$. If $y \neq y_s$, it indicates that the key information is contained in the clause, and add s_i as a candidate sentence to S' . Then we use jieba library (A Chinese word segmentation package of python) to tag and record the part-of-speech of all words X in the obtained candidate sentences as a “word: part-of-speech” dictionary. We remove the words with meaningless part-of-speech $POS = \{prep., pron., num., art.\}$ in X to obtain the candidate keywords X' .

C. WORD CONTRIBUTION CALCULATION ALGORITHM

In text classification tasks, different words may have different sentiment classification tendencies. To modify fewer words but change the text tendency most, finding the words that have the largest contribution to the original category is the key operation of the algorithm. A word with a high contribution

means that the ability to classify into the current category will be greatly reduced after removing the word. We rank and locate candidate words according to the impact on category contributions. The contribution of each word is measured by the following methods:

$$C_F(x_i, y_i) = F(x_1, x_2, \dots, x_i, \dots, x_n) - F(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

To get a quantitative representation of the contribution value, the confidence degree P is introduced to calculate specifically:

$$C_F(x_i, y_i) = P_F(y_i | S') - P_F(y_i | S'_{\tilde{x}_i})$$

where $P_F(y_i | S')$ is the probability that the text gets the predicted label y_i according to the classifier F , $S'_{\tilde{x}_i}$ represents the text after deleting the word x_i . For a long piece of text after “purification” operation, the text length and the time to calculate the contribution value $C_F(x_i, y_i)$ will be reduced so that can better determine the contribution of each word for the particular classifier F .

D. KEYWORD MODIFICATION STRATEGY

The key to generating adversarial text is to add perturbation on certain words x in the sentence S so that makes the generated text S' does not affect human normal reading but fool a text detector or classifier. According to current research, keyword modification strategies in English for adversarial texts can be summarized from reference [23], [25], and [29] as follows: (1) replace original words with synonyms; (2) randomly exchange adjacent letters in words; (3) replace a certain letter in a word with other characters; (4) randomly insert letters in a word; (5) randomly delete letters other than the first and last letters in a word, etc. However, the above method can't apply to Chinese text as the basic unit of English is 26 letters that most of them have no practical meanings, and the modification of individual letters does not affect the semantics of words. The basic unit of Chinese is the thousands of Chinese characters commonly used that always express different semantics. Therefore, the word modification strategy based on Chinese characters requires diversified attempts and strategic choices. Based on the above analysis, we attempt to use three Chinese keyword modification strategies to generate adversarial samples that achieve the purpose of fooling the deep neural networks with small changes compared to the original text. Examples of the modification strategies are shown in Table 1:

TABLE 1. Examples of keyword modification strategies.

Original words	CCE	CI	CCSR
清白 (Innocent)	白洁	清白→	𠄎吉白
评论家 (Critic)	论评家	评^论家	讠平讠仑佳
数据结构 (Data Structure)	数结构	数据结」构	娄女才居屮吉木勾

1) CHINESE CHARACTER EXCHANGE (CCE)

Exchange the position of Chinese characters in the words. Although the change of the position of Chinese characters seems not to guarantee semantic Continuity theoretically, psychological studies [34] have shown that humans can read and understand the scrambled text, because the reading inertial thinking will automatically complete the ordering of text to understand the purpose of semantics.

2) CHARACTER INSERTION (CI)

Randomly insert disturbing symbols in words. Artificially create a set of disturbing symbols, which is composed of symbols that have no practical meaning and do not affect the semantics of the text, such as punctuation marks, Roman characters, etc.

3) CHINESE CHARACTER SPLIT AND REPLACEMENT (CCSR)

Chinese characters can be divided into upper and lower structures or left and right structures. Due to the way humans read from left to right, the left-right split text is only slightly different for human observers than the original text. Therefore, we propose a method of Chinese characters split and replacement that uses split variants to replace the characters with left-right structure and then use homophones to replace the other characters. Although the glyphs of Chinese characters have changed, humans can still accurately grasp the semantics of sentences through context. CCSR first constructs a dictionary manually which contains all Chinese characters with left-right structure and the split Chinese character variants. The original text is replaced with the variants by comparing the text with the dictionary. Meanwhile, another dictionary of homophones is constructed manually to ensure that every Chinese character can find a homophone that can be replaced.

E. GENERAL DESCRIPTION OF THE ALGORITHM

Based on the above-mentioned word contribution value calculation algorithm and keyword modification strategy, we propose a text adversarial sample generation method for Chinese characters. First, we perform word segmentation on the text and then divide the text into clauses to obtain the clause set S_{seg} ; delete each clause s_i of the original text in turn, and scrutinize whether the predicted label is the same as the original label. If it is different, add s_i into candidate key sentence set S' ; Secondly, tag the clauses in the candidate sentence set and delete the meaningless part of speech $POS = \{prep., pron., num., art.\}$ to get the candidate keyword set X' . Calculate the contribution scores C of each keyword in descending order. Finally, we take the keyword modification strategy function $T(\cdot)$ to modify the keywords and predict the labels respectively. σ is the set maximum modification threshold that within the threshold range the operation amplitude changes dynamically. If the predicted label changes, an adversarial example is successfully

TABLE 2. Experimental datasets.

Item	Ctrip Hotel Reviews dataset	JD.com product review dataset
Training Data	12000	4800
Test Data	3000	3000
Average Length	139	43
Median Length	90	26

generated and we no longer modify the text. $T(\cdot)$ can be any of the three keyword modification methods, and $Cost(\cdot)$ is the cumulative frequency of text modification.

The WordChange algorithm is described as follows:

Algorithm 1 WordChange

Input: Text S ; Text Category Label y ; RNN Classifier $F(\cdot)$; Modification Strategy Function $T(\cdot)$; Operation Threshold σ

Output: Adversarial Text \hat{S}

```

1:  $S_{seg} = \{s_1, s_2, s_3, \dots, s_n\}$  ( $seg = ", \circ?!"$ )
2: for  $i = 1, \dots, n$  do
3:    $y_i = F(S_{seg} - s_i)$ 
4:   if  $y_i = y$  then
5:      $S' \leftarrow \text{Add } s_i \text{ into } S'$ 
6:   end if
7: end for
8: Fill the part-of-speech  $POS = \{\text{prep.}, \text{pron.}, \text{num.}, \text{art.}\}$ .
9:  $X' = \{x_1, x_2, x_3, x_4, \dots, x_n\}$ 
10: for  $i = 1, \dots, n$  do
11:    $C_F(x_i, y_i) = P_F(y_i|S') - P_F(y_i|S'_{\hat{x}_i})$ 
12: end for
13:  $X'_{sorted} \leftarrow \text{Sort } C_F(x_i) \text{ by descending } C_F(x_i)$ 
14: for  $x_i$  in  $X'_{sorted}$  do
15:    $\hat{x}_i = T(x_i)$ 
16:   while  $Cost(x_i, \hat{x}_i) < \sigma$  do
17:      $x_i = \hat{x}_i$ 
18:     if  $F(\hat{S}) \neq y$ 
19:       return  $\hat{S}$ 
20:     end if
21:   end while
22: end for
23: return  $\hat{S}$ 

```

IV. EXPERIMENTAL RESULTS AND ANALYSIS OF TEXT SENTIMENT CLASSIFICATION

Sentiment analysis is also called opinion mining that is the process of analyzing, processing, inducing, and inferring subjective texts with emotional color [35]. Sentiment analysis text is a kind of subjective text with emotion, including human attitudes and opinions on entities such as products, services, organizations, etc. Potential users can browse the commentary text to understand the views of the public. In this section, we evaluate the effectiveness of adversarial text generated from sentiment analysis datasets. Firstly we introduce the

experimental datasets, models, baseline methods and evaluation criteria; then evaluate the experimental results and analyze the effectiveness of the proposed method; finally, transfer the generated adversarial text to the Chinese sentiment analysis platform to observe the transfer performance.

A. EXPERIMENTAL SETUP

We use two public benchmark datasets as the experimental data for the adversarial sample of sentiment analysis: Ctrip Hotel Reviews dataset¹ and JD.com product review dataset¹. Both sets of data use 1 and -1 to represent positive and negative samples. The specific dataset information is shown in Table 2.

To evaluate the effectiveness of the method intuitively, a Word-LSTM (Long Short-Term Memory Network) [16] model is used as the text-based attack target. Because the LSTM model has a good performance on natural language processing tasks and can measure the effectiveness of our method better. The network contains a random embedding layer to accept word input. The embedding vectors are then fed through five LSTM layers where each layer has 100 hidden nodes. The hidden state of LSTM layers is fed to the fully connected layer with a LogSoftMax activation function to get the final classification confidence value. We set the learning rate to 0.0005, the batch size to 128, and the maximum number of epochs to 20 during training. In deep learning modeling, the unknown words will be mapped to the "unknown" embedding vector. The maximum modification threshold is set to 30. Attack performance is measured by the accuracy of classification. The lower classification accuracy of the model, more effective the attack method is.

B. EXPERIMENTAL RESULTS

1) BASIC EXPERIMENTS

Table 3 summarizes the experimental results and performance compared with WordHandling[33] and DeepWordBug[29].

The keyword contribution value calculation algorithm and three different modification strategies proposed by our method have achieved good attack results on two sets of data sets, and the effect is better than the baselines. The CCE strategy can achieve an average decrease of 32.94%, the CI strategy can achieve an average decrease of 44.41%, and the CCSR can reduce the classification accuracy by 45.44%. In summary, the WordChange method can effectively

¹<https://github.com/cgq666/Chinese-text-sentiment-classification-dataset>

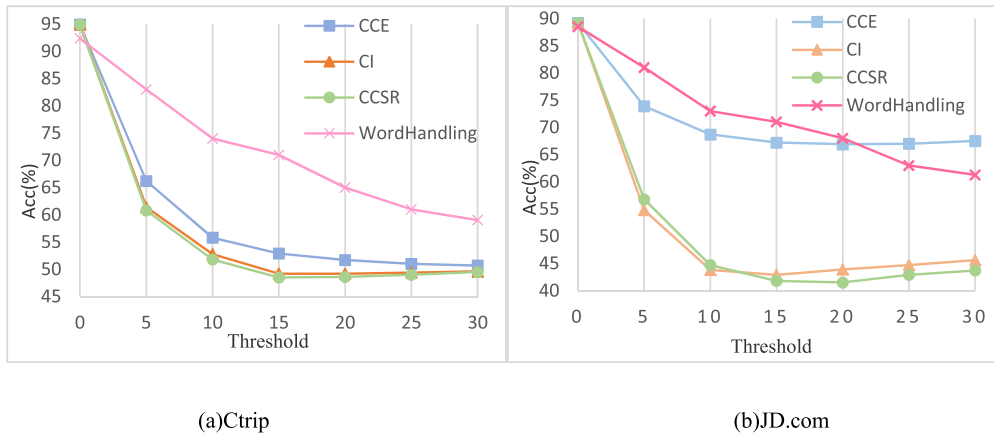


FIGURE 2. Accuracy of adversarial examples with operating threshold on sentiment analysis datasets.

TABLE 3. Experimental results of sentiment analysis datasets.

Method \ Dataset	Original	WordHandling[33]		DeepWordBug[29]		WordChange					
						Modification Strategies					
						CCE		CI		CCSR	
Acc(%)	Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease	
Ctrip	94.93	62.19	32.74	71.07	23.86	50.77	44.16	49.66	45.27	49.53	45.40
JD.com	89.21	62.89	26.32	66.76	22.45	67.50	21.71	45.66	43.55	43.73	45.48

generate adversarial text with great performance, and use three keyword modification strategies to implement a variety of attacks.

2) THRESHOLD ANALYSIS

The operation threshold σ is a dynamic parameter that represents the maximum number of keywords modifications. To explore the impact of thresholds on the utility of the generated adversarial text, experiments were performed on different thresholds. We take the same experimental conditions and parameters as WordHandling, select 1000 pieces of data longer than 120 words, and the maximum modification range is also set to 30. The accuracy of adversarial examples with different operating thresholds on sentiment analysis datasets is summarized in Figure 2. With the increase of the threshold, the text-modifiable operating space continues to increase. When the threshold reaches 15, the model accuracy becomes stable. It proved that our method can be used in a smaller operating space than WordHandling and achieve more effective attacks.

3) ADVERSARIAL SAMPLE QUALITY

To measure the quality of the adversarial samples generated by WordChange, Word Mover’s Distance (WMD) [36] method was used to test the similarity between the generated text and the original text. The smaller the WMD score, the higher the similarity between the texts. In the three

modification strategies, 2000 pieces of data were randomly selected for testing, and we set the same experimental conditions as WordHandling. Table 4 shows the proportion of data in each interval of the WMD score. The score occupies the largest proportion in the 0-0.2 interval, which verifies that the sample generated by WordChange has higher quality. Note that the adversarial text generated by the CCSR modification strategy in Table 3 has the best attack performance. However, the CCSR method has the worst adversarial sample quality in Table 4. This is because CCSR method is slightly stronger in modifying words than CCE and CI methods, so the quality of the text is not as good as the other two methods.

C. DISCUSSION AND ANALYSIS

1) IMPACT OF DIFFERENT WORD SEGMENTATION METHODS

Word segmentation refers to the process of recombining consecutive sequences into word sequences by certain specifications. In English, spaces are used as natural delimiters between words. Sentences and paragraphs can be easily separated by obvious delimiters but words do not have a formal delimiter in Chinese. The Chinese word segmentation is much more complex and difficult than English, and it has gradually become a research hotspot. We research the impact of the common word segmentation methods such as jieba, THULAC [37], and FoolNLTK on the generation of adversarial texts. Due to different specific word segmentation

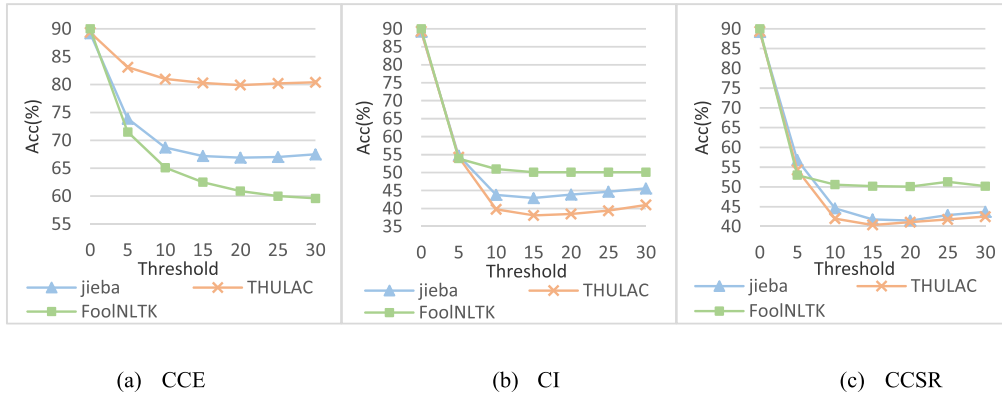


FIGURE 3. The classification accuracy of Chinese word segmentation methods on JD.com dataset.

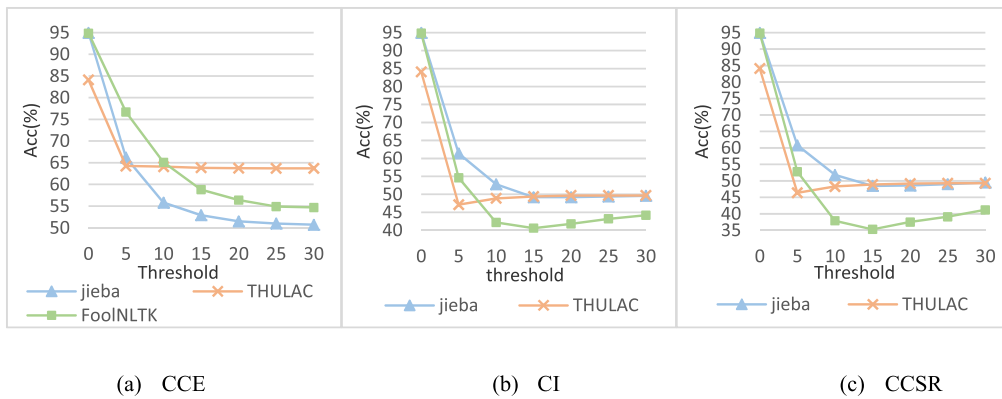


FIGURE 4. The classification accuracy of Chinese word segmentation methods on ctrip dataset.

TABLE 4. Proportion of sample numbers to total samples in WMD intervals.

Range	WordHanding[33]	DeepWordBug [29]	WordChange		
			CCE	CI	CCSR
0-0.2	0.5%	0.2%	93.4%	99.9%	35.2%
0.2-0.4	14.9%	6.0%	6.4%	0.06%	21.6%
0.4-0.6	35.8%	25.7%	0.1%	0.04%	17.9%
0.6-0.8	18.7%	32.4%	0.06%	0%	5.5%
0.8-1.0	16.3%	23.3%	0.03%	0%	6.7%
1.0and above	13.8%	12.4%	0.01%	0%	13.1%

algorithms, the word segmentation results of the same sentence are different. Table 5 shows examples of different word segmentation results.

Different word segmentation strategies may also have an impact on the generation of adversarial samples. We generate adversarial text for the aboveword segmentation methods and explores the difference in their ability to deceive classification models. The experimental results are shown in Figure 3 and Figure 4. The adversarial text generated by different word segmentation methods effectively reduced the accuracy of the classifier, and the accuracy difference among them is not

large, which illustrates that our attack method can be applied to multiple word segmentation strategies.

2) ANALYSIS OF MODIFICATION STRATEGIES

We also explore the performance of several modification strategies, namely homophone replacement strategy (HR), Chinese character splitting (CCS), and Tongue-flatted or Tongue-rolled Pronunciationreplacement (TTPR). This part of the experiment is to explore the diversity of Chinese adversarial text generation strategies, but also provides more ideas for future defense work. Figure 5 summarizes the

TABLE 5. Examples of different Chinese word segmentation methods.

Original Sentence: 还行,穿着舒服,鞋底也柔软透气,不会累脚,下次就选这家店买的。
(It's okay, comfortable to wear, and the soles are soft and breathable, so it won't tire your feet. I'll choose this store next time.)
jieba: 还行 , 穿着舒服 , 鞋底也柔软透气 , 不会累脚 , 下次就选这家店买的 .
THULAC: 还行 , 穿着舒服 , 鞋底也柔软透气 , 不会累脚 , 下次就选这家店买的 .
FoolNLTK: 还行 , 穿着舒服 , 鞋底也柔软透气 , 不会累脚 , 下次就选这家店买的 .

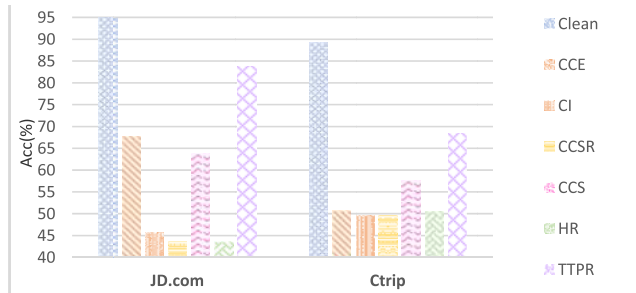


FIGURE 5. Experimental results for multiple modification strategies.

experimental results for all modification strategies on sentiment analysis datasets.

- HR:** The homophonereplacement strategy is a modification strategy used in WordHandling which means two Chinese characters have the same pinyin code. We expand the homophone replacement dictionary to a certain extent and almost cover all Chinese with homophones. The experimental results show that the model effect can be reduced by 44.9% on average.

- CCS:** Since there are not many detachable Chinese characters with left-right structures, we combine Chinese character splitting(CCS) and homophonereplacement(HR) as Chinese character split and replacement(CCSR) strategy above, which can avoid the situation where the targeted keywords cannot be completely modified and reduce the unreadable text replaced by too many homophones. Although the splitting method may not be able to modify all the keywords like other strategies, the classification performance of the model can still reduce the average performance of the model by 31.36%

- TTPR:** Tongue-flatted or Tongue-rolledPronunciation is a unique characteristic of Chinese characters. The so called tongue-flatted pronunciation refers to issue the *z, c, s* (pinyin code) that the tongue protrudes flatly against or near the upper teeth. The tongue-rolledpronunciation refers to the tip of the tongue rising, touching or approaching the front hard palate, and issue the *zh, ch, sh* and *r* (pinyin code). Tongue-flatted and tongue-rolledare issued different from each other but sound similar. Inspired by this, the replacement of tongue-flatted or tongue-rolled pronunciation is also understandable through pronunciation association and contextphrase. TTRP does not modify the keywords comprehensivelycause Chinese characters have a limited number of tongue-flatted or tongue-rolled pronunciation, but it also has a certain attack performance that reduces the classification accuracy by 16%.

3) TRANSFERABILITY

The adversarial samples generated for one classification model can also successfully fool other classification models with the same task, indicating that the adversarial samples are transferable. In the field of computer vision, Papernot *et al.* [38] have confirmed that generating adversarial examplesby producingwhite-box attacks on an alternative model, an effective black-box attack can be implemented on the target model. In the natural language domain, the transferability of Chinese adversarial texts is also effective.

To investigate whether the Chinese adversarial text has this attribute, this article saves the adversarial text generated on the LSTM [16] model and evaluates their effect on other models/platforms. Due to the results in the threshold analysis experiment, a better attack effect and smaller text modifications can be obtained when the threshold was 15. Therefore, we set the experimental operation threshold to 15. We applied two deep learning classification networks, TextCNN [39] and DPCNN [40], as the models to which our generated adversarial samples transfered. During the training of the two networks, the learning rate was 0.001, the batch size was 64, and the maximum epoch was 50. We also added two Chinese sentiment analysis APIs (Baidu AI https://ai.baidu.com/tech/nlp/sentiment_classify sentiment platform and Tencent Cloud² sentiment analysis platform) as migration platforms. The results are shown in Table 6.

As observed in Table 6, the accuracy of the classification results is all reduced in the transferability evaluation of two datasets. Most adversarial texts can be successfully migrated to other models or even text detection platforms. For example, the adversarial texts generated by the Ctrip dataset have a success rate of 66.55% when attacking the DPCNN model, and the original accuracy rate is above 96%. The reduction in classification accuracy can reach a maximum of 34.75% on DPCNN model. Consequently, the adversarial text generated by WordChange can successfully implement adversarial attacks across multiple models and platforms. In particular, for the services provided by Tencent Cloud, the CI strategy cannot completely reduce its classification accuracy. We guess that the service will filter out all the useless special characters in Chinese text when preprocessing the input data. Overall, the CCSR strategy has the best transferability performance, which illustrates that the adversarial

² <https://cloud.tencent.com/product/nlp>

TABLE 6. Transferability of adversarial examples on sentiment analysis datasets.

Dataset	OriginalAcc(%)	Model/Platform	WordChange					
			CCE		CI		CCSR	
			Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease
Ctrip	80.93	TextCNN[39]	53.20	27.73	58.40	22.53	55.46	25.47
	96.36	DPCNN[40]	88.29	8.07	88.69	7.67	66.55	29.81
	88.39	Baidu AI	79.80	8.59	85.56	2.83	63.73	24.66
	85.46	Tencent Cloud	81.1	4.36	85.51	-0.05	59.78	25.68
JD.com	85.16	TextCNN[39]	66.40	18.76	59.30	25.86	51.26	33.90
	92.30	DPCNN[40]	88.76	3.54	88.56	3.74	57.55	34.75
	93.37	Baidu AI	88.35	5.02	91.72	1.65	63.5	29.87
	88.84	Tencent Cloud	86.68	2.16	88.91	-0.07	61.52	27.32

TABLE 7. Comparison of attacks on TextCNN and LSTM models.

Model	Dataset	Original	WordChange					
			CCE		CI		CCSR	
			Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease
TextCNN[39]	Ctrip	85.16	49.47	35.46	47.70	37.23	47.57	37.36
	JD.com	80.93	52.61	28.32	48.69	32.24	48.61	32.32
LSTM[16]	Ctrip	94.93	50.77	44.16	49.66	45.27	49.53	45.40
	JD.com	89.21	67.50	21.71	45.66	43.55	43.73	45.48

examples generated by focusing on features of Chinese can achieve more effective attacks.

4) COMPARISON OF ATTACKS ON OTHER TEXT CLASSIFICATION MODELS

In the experimental setup, we take into account the better classification performance of the LSTM model, and adversarial attacks on it can effectively evaluate our method. Therefore, the experiments are all performed on LSTM models in this paper. In this section, we further verify the effectiveness of our method on TextCNN [39]. We used the same dataset to train the TextCNN network and get a pre-trained model. During training, the learning rate was 0.001, the batch size was 64, and the maximum epoch was 50. Table 7 shows the comparison of the experimental results on TextCNN [39] and LSTM [16]. As observed in Table 7, the attack performance on the TextCNN model is slightly less than that of the LSTM model. We think it is because the original classification accuracy of the TextCNN model is relatively low. In summary, our method can effectively attack LSTM model as well as TextCNN model.

V. EXPERIMENT RESULTS AND ANALYSIS OF SPAM DETECTION

Spams can easily contain some false information (advertising, financing promotion, gambling information, etc.). When an attacker adds a counter sample to the email, it will cause

TABLE 8. Spam dataset.

Item	Spam Dataset
Training Data	6999
Test Data	3002
Average Length	143
Median Length	156

the detection system to incorrectly divide spams into normal emails or classify normal mails as spams, which will increase the probability of users clicking on virus-carrying emails and affect network security. Exploring the security issues against spam adversarial samples effectively promotes the robustness of deep models and can also evaluate the universality of our method comprehensively. In this section, we mainly show the performance of adversarial text generated on the spam dataset.

A. EXPERIMENTAL SETUP

We use a public spam corpus consisted of an English dataset (trec06p) and a Chinese dataset (trec06c)³ from the International Text Retrieval Conference. The trec06c was cleaned and the encoding format was converted to utf-8 format as the experimental dataset, the specific information is shown in Table 8.

³<https://plg.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/foo06>

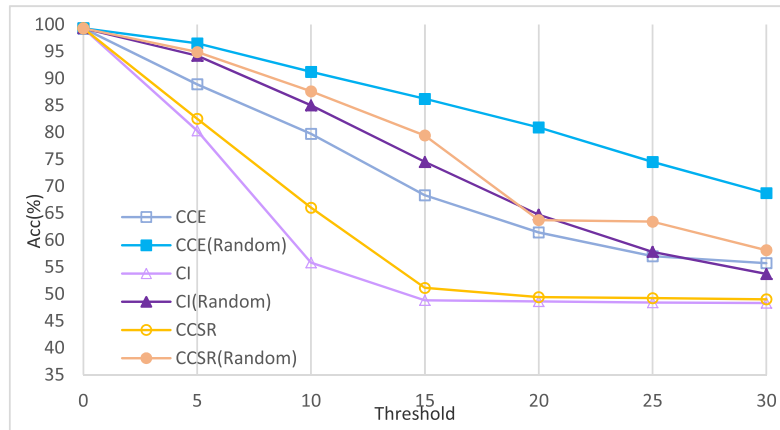


FIGURE 6. Accuracy of adversarial examples with operating threshold on spam dataset.

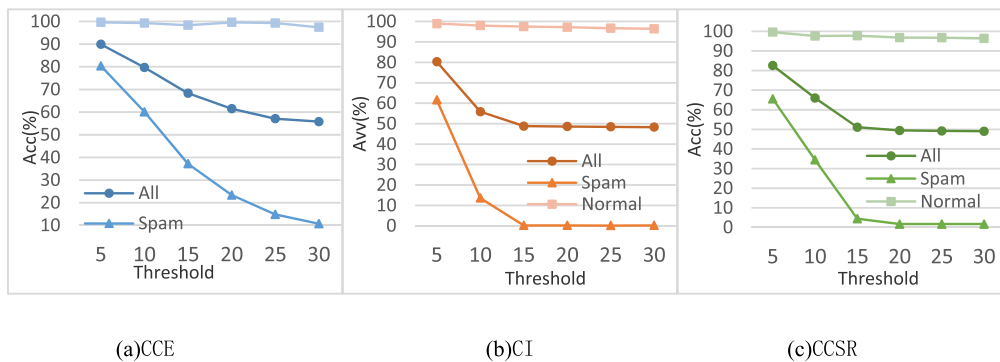


FIGURE 7. Impacts of spam/ham dataset on adversarial examples.

We take a total of selected 10001 data with spams and normal emails as samples. The spam category is marked as -1 , and the normal email category is marked as 1 . The target models are the same as in Chapter 4. The attack performance is measured by the accuracy rate of spam detection, that is, the spam is not consistent with the actual label of the original email, indicating that the method can successfully spoof spam detection systems to achieve attacks.

B. EXPERIMENT RESULTS

We take a randomly choice of words as a benchmark method for comparison. Meanwhile, TF-IDF [41] and TextRank [42] are also used as the benchmark keyword selection algorithm. Table 9 summarizes the results of the attack on the LSTM [16] model and the performance from different modification strategies.

As observed in Table 9, a high attack success rate can also be achieved on spam detection. Compared with the baseline methods, we can intuitively observe the experimental results of our keyword contribution value algorithm on different modification strategies that demonstrate the superiority of the proposed approach. Figure 6 shows the effect of different operation threshold σ on the performance of adversarial text. The accuracy of the spam dataset gradually decreases as the operation threshold increases. With a threshold of 30, the adversarial text has the highest fool rate.

C. DISCUSSION AND ANALYSIS

Through the analysis of the above experimental results, it can be initially observed that the spam dataset and the sentiment analysis dataset have similar experimental results. However, during the experiment, the performance of the positive and negative samples on the spam dataset is particularly different. To explore this issue, we evaluate the success rate of 1500 positive and negative samples separately. Figure 7 shows the results of the three modification strategies on the normal mails and the spams.

The accuracy rate of spams gradually decreases to less than 10%, while the normal email remains at 90%. We believe that: the content of spam is mostly commercial advertising, porn marketing, scams or phishing sites. The feature of the spams is relatively singular and concentrated, while the content of the normal email is more extensive and diverse.

To further explore the reasons, we randomly select 3,000 spams and 3,000 normal emails from the training data and testing data, and they were made into a word cloud and observed the keywords that appear more frequently in the text, as shown in Figure 8. The results show that the high-frequency words of spams are concentrated in “电子科技” (electronic technology), “国际” (international), “服务” (service), “热线” (hotline), etc., and the information of normal email is more discrete and common. After adding disturbance to the keywords of the spam, the



FIGURE 8. Word cloud distribution of the spam dataset.

TABLE 9. Experimental results of spam dataset.

Method	Original	CCE		CI		CCSR	
	Acc(%)	Acc(%)	Decrease	Acc(%)	Decrease	Acc(%)	Decrease
Random	99.33	68.75	30.58	53.79	45.54	58.12	41.21
TF-IDF[41]		68.30	31.03	54.27	45.06	57.14	42.19
TextRank[42]		64.53	34.8	52.74	46.59	54.37	44.96
WordChange		55.79	43.54	48.32	51.01	49.06	50.27

remaining text will get a higher positive score, while normal email still has many normal texts which are insufficient to obtain a higher score for predicting to be spam.

VI. CONCLUSION

In this paper, we propose a Chinese adversarial text generation strategy based on multiple modification strategies named WordChange. It is efficiently and accurately misleading the classification model under the black-box condition of unknown model details to deceive security systems. Our method first implements the text filtering operation and filters words with no actual semantics to form a candidate keyword pool; then uses the keyword contribution score to calculate the importance of the words. The approach of extracting keywords based on clauses can effectively reduce the search space and locate words more accurately. Meanwhile, we introduce Chinese character exchange strategies based on reading inertia thinking; character insertion strategy with adding disturbed symbols; and Chinese character split and replacement strategy based on glyph structure and pinyin characteristics. The experimental results show that WordChange can generate better and higher quality adversarial samples on both the sentiment analysis dataset and the spam dataset. The average classification accuracy of the LSTM [16] model is reduced by 45% and 48%. We also evaluate the effect of the adversarial samples based on multiple word segmentation processes which proves that our method is versatile. Besides, we expand more modifiable operations for Chinese text, such as Tongue-flatted or Tongue-rolled Pronunciation replacement, homophone replacement, etc. For other text classification models or online platforms, the transferable of adversarial samples also implies that they have vulnerabilities that can be attacked. We hope our study will

provide more ideas and possibilities for further research on deep neural networks and Chinese natural language processing.

REFERENCES

- [1] C. Szegedy et al., “Intriguing properties of neural networks,” in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.
- [2] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Proc. ICLR Workshop*, 2017, pp. 1–14.
- [3] Y. Liu, W. Zhang, and N. Yu, “Protecting privacy in shared photos via adversarial examples based stealth,” *Secur. Commun. Netw.*, vol. 2017, pp. 1–15, Nov. 2017.
- [4] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, “Attacking visual language grounding with adversarial examples: A case study on neural image captioning,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, Jul. 2018, pp. 2587–2597.
- [5] Y. Zhang, H. Gao, G. Pei, S. Kang, and X. Zhou, “Effect of adversarial examples on the robustness of CAPTCHA,” in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2018, pp. 1–10.
- [6] Y. Wu, D. Bamman, and S. Russell, “Adversarial training for relation extraction,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1778–1783.
- [7] J. Ebrahimi, D. Lowd, and D. Dou, “On adversarial examples for character-level neural machine translation,” in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, Santa Fe, NM, USA, Aug. 2018, pp. 653–663.
- [8] T. Niu and M. Bansal, “Adversarial over-sensitivity and over-stability strategies for dialogue models,” in *Proc. 22nd Conf. Comput. Natural Lang. Learn. (CoNLL)*, Brussels, Belgium, Oct. 2018, pp. 486–496.
- [9] P. Minervini and S. Riedel, “Adversarially regularising neural NLI models to integrate logical background knowledge,” in *Proc. 22nd Conf. Comput. Natural Lang. Learn. (CoNLL)*, Brussels, Belgium, Oct. 2018, pp. 65–74.
- [10] K. Wang and R. Xia, “Adversarially regularising neural NLI models to integrate logical background knowledge,” in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*, Brussels, Belgium, Oct. 2018, pp. 65–74.
- [11] G. Chang and H. Huo, “A method of fine-grained short text sentiment analysis based on machine learning,” *Neural Netw. World*, vol. 28, no. 4, pp. 325–344, 2018.
- [12] P. Liu, H. Zhao, J. Teng, Y. Yang, Y. Liu, and Z. Zhu, “Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark,” *J. Central South Univ.*, vol. 26, no. 1, pp. 1–12, 2019.

- [13] M. Zhang, X. Ai, and Y. Hu, "Chinese text classification system on regulatory information based on SVM," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 252, Jul. 2019, Art. no. 022133.
- [14] F. Huang, S. Zhang, J. Zhang, and G. Yu, "Multimodal learning for topic sentiment analysis in microblogging," *Neurocomputing*, vol. 253, pp. 144–153, Aug. 2017.
- [15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] J. Wenzhen, Z. Hong, and Y. Guocai, "An efficient character-level and word-level feature fusion method for Chinese text classification," *J. Phys., Conf. Ser.*, vol. 1229, May 2019, Art. no. 012057.
- [17] M. Zhu and X. Yang, "Chinese texts classification system," in *Proc. IEEE 2nd Int. Conf. Inf. Comput. Technol. (ICICT)*, Mar. 2019, pp. 149–152.
- [18] H. Tao, S. Tong, H. Zhao, T. Xu, B. Jin, and Q. Liu, "A radical-aware attention-based model for Chinese text classification," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 5125–5132.
- [19] X. Qiao, C. Peng, Z. Liu, and Y. Hu, "Word-character attention model for Chinese text classification," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 12, pp. 3521–3537, Dec. 2019.
- [20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.
- [21] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.
- [22] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, May 2017, pp. 39–57.
- [23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
- [24] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Sep. 2017, pp. 2021–2031.
- [25] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4208–4215.
- [26] S. Samanta and S. Mehta, "Generating adversarial text samples," in *Proc. 40th Eur. Conf. IR Res. (ECIR)*, Grenoble, France, Mar. 2018, pp. 744–749.
- [27] Z. Gong, W. Wang, B. Li, D. Song, and W.-S. Ku, "Adversarial texts with gradient methods," 2018, *arXiv:1801.07175*. [Online]. Available: <https://arxiv.org/abs/1801.07175>
- [28] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-box adversarial examples for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, Jul. 2018, pp. 31–36.
- [29] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, San Francisco, CA, USA, May 2018, pp. 50–56.
- [30] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "TextBugger: Generating adversarial text against real-world applications," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Feb. 2019, pp. 1–15.
- [31] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul./Aug. 2019, pp. 1085–1097.
- [32] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 1875–1885.
- [33] W. Q. Wang, R. Wang, L. N. Wang, and B. X. Tang, "Adversarial examples generation approach for tendency classification on Chinese texts," (in Chinese), *Ruan Jian Xue Bao/J. Softw.*, vol. 30, no. 8, pp. 2415–2427, 2019. [Online]. Available: <http://www.jos.org.cn/1000-9825/19/1565.htm>
- [34] M. Davis. (2003). *Aoccdnrig to a Rscheearch at Cmabrigde Uinervtisy*. [Online]. Available: <http://www.mrcbu.cam.ac.uk/people/matt.davis/cmabridge/>.
- [35] Y.-Y. Zhao, B. Qin, and T. Liu, "Sentiment analysis," *Ruan Jian Xue Bao/J. Softw.*, vol. 21, no. 8, pp. 1834–1848, 2010.
- [36] M. J. Kusner, Y. Sun, N. I. Kolkun, and K. Q. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.
- [37] M. Sun, X. Chen, K. Zhang, Z. Guo, and Z. Liu, "THULAC: An efficient lexical analyzer for Chinese," Tech. Rep., 2016.
- [38] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (Asia CCS)*, Apr. 2017, pp. 506–519.
- [39] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–6.
- [40] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 562–570.
- [41] G. Salton and C. T. Yu, "On the construction of effective vocabularies for information retrieval," *ACM SIGIR Forum*, vol. 9, no. 3, pp. 48–60, Dec. 1974.
- [42] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2004, pp. 404–411.



NUO CHENG is currently pursuing the M.S. degree with the School of Cyber Engineering, Xidian University, Xi'an, Shaanxi, China. Her current research interests include Captcha, computer security, and machine learning.



GUO-QIN CHANG is currently pursuing the Ph.D. degree with the School of Cyber Engineering, Xidian University, China. Her current research interests include cyber engineering, computer security, and deep learning.



HAICHANG GAO (Member, IEEE) received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2006. He is currently a Professor at the School of Computer Science and Technology, Xidian University. He has published more than 30 articles. He is currently in charge of a project of the National Natural Science Foundation of China. His current research interests include Captcha, computer security, and machine learning.



GE PEI is currently pursuing the M.S. degree with the School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China. Her current research interests include Captcha, computer security, and machine learning.



YANG ZHANG received the M.S. degree from the School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China, in 2018. Her current research interests include Captcha, computer security, and machine learning.

...