IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 July 2024, accepted 25 July 2024, date of publication 29 July 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435573

RESEARCH ARTICLE

IAE: Irony-Based Adversarial Examples for Sentiment Analysis Systems

XIAOYIN YI[1,2] **AND JIACHENG HUANG**[2]

[1]Chongqing Key Laboratory of Public Big Data Security Technology, Chongqing 400000, China
[2]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Jiacheng Huang (Dylan.JiaCheng.Huang@outlook.com)

This work was supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJQN202302403.

ABSTRACT Adversarial examples, which are inputs deliberately perturbed with imperceptible changes to induce model errors, have raised serious concerns for the reliability and security of deep neural networks (DNNs). While adversarial attacks have been extensively studied in continuous data domains such as images, the discrete nature of text presents unique challenges. In this paper, we propose Irony-based Adversarial Examples (IAE), a method that transforms straightforward sentences into ironic ones to create adversarial text. This approach exploits the rhetorical device of irony, where the intended meaning is opposite to the literal interpretation, requiring a deeper understanding of context to detect. The IAE method is particularly challenging due to the need to accurately locate evaluation words, substitute them with appropriate collocations, and expand the text with suitable ironic elements while maintaining semantic coherence. Our research makes the following key contributions: (1) We introduce IAE, a strategy for generating textual adversarial examples using irony. This method does not rely on pre-existing irony corpora, making it a versatile tool for creating adversarial text in various NLP tasks. (2) We demonstrate that the performance of several state-of-the-art deep learning models on sentiment analysis tasks significantly deteriorates when subjected to IAE attacks. This finding underscores the susceptibility of current NLP systems to adversarial manipulation through irony. (3) We compare the impact of IAE on human judgment versus NLP systems, revealing that humans are less susceptible to the effects of irony in text.

INDEX TERMS Adversarial examples, sentiment analysis, irony-based, black-box.

I. INTRODUCTION

Adversarial examples [1], crafted by adding imperceptible tiny perturbations to origin inputs maliciously, cause deep neural networks (DNNs) to fail blatantly. The secure issue, namely adversarial attack, is being widely concerned among researchers as soon as it was proposed. Extensive research has revealed that adversarial examples widely exist in many fields, e.g., computer vision (CV) [2], natural language processing (NLP) [3] and automatic speech recognition (ASR) [4].

Textual data is not as continuous as images which are capable of being perturbed imperceptibly with pixel noise. Instead, it is impossible to craft a factual imperceptible perturbation on a text due to its discrete nature. Furthermore, the grammar and semantics may be broken easily by changing even a character. The textual adversarial attack is confronted with greater challenges compared with images.

A variety of textual adversarial attack models has been proposed in many NLP tasks, incorporating machine translation [5], question-answering system [3], sentiment analysis [6], et al. Spelling mistake [7], visually similar characters substitution [8], synonyms substitution [9] and sentence paraphrasing [10] are typical textual adversarial attack methods ranging from word-level to sentence-level while categorized by attacking granularity. However, there are still a few issues while assuming those methods in practical situations: 1) Subtle spelling mistakes can be

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

TABLE 1. Examples of straightforward and ironic text.

Straightforward 1: 他真是个糟糕的守门员，让对方进了六个球。He is a really terrible goalkeeper, allowing the other side to score six goals.
Ironic 1: 他真是个有天赋的守门员，让对方进了六个球。 He is a really talented goalkeeper, allowing the other side to score six goals.
Straightforward 2: 那个男人真恶心，在公共场所随地吐痰。 That man is totally disgusting, spiting everywhere in public.
Ironic 2(a): 那个男人真美味，在共场所随地吐痰。 That man is totally delicious, spiting everywhere in public.
Ironic 2(b): 那个男人真优雅，在共场所随地吐痰。 That man is totally elegant, spiting everywhere in public.
Ironic 2(c): 那个男人真优雅，在共场所随地吐痰。真是值得称赞啊。 That man is totally elegant, spiting everywhere in public. It is really praiseworthy.

recovered easily with spelling error correction [11]. 2) Words out of vocabulary may arise attention and alertness while exceeding averages in a text. 3) Word substitution and sentence paraphrasing may cause grammar to be broken or semantics deviated. Therefore, we consider a textual adversarial attacking method more practically.

The irony is a kind of rhetorical device expressing a strong emotion referring to the opposite of literal meaning and needs to understand the actual meaning from context. Detecting irony is challenging while implementing it the model needs to have human-level language understanding ability. As far as we know, there are no studies considering converting text from straightforward to ironic as a method of generating textual adversarial examples orienting the NLP task of sentiment analysis presently.

The cruxes of converting a text from straightforward into ironic are to turn the polarity of the evaluation words and make an ironic expansion appropriately when necessary. Specifically, there are at least three challenges here: 1) locating evaluation words, 2) substituting evaluation words with correct collocation, and 3) expanding text with appropriate ironic evaluation.

Without loss of generality, we consider Chinese irony-based adversarial examples in this paper. As shown in Table 1, Chinese words "糟糕" is an evaluation to "守门员" in first sentence, where "糟糕" means "*terrible*" and "守门员" means "*goalkeeper*". It is necessary to locate the words "糟糕" as an evaluation disclosing negative emotion and then substitute "糟糕" with "有天赋" which means "*talented*". Humans are in capable of understanding the second sentence still exhibiting negative emotion with strong language comprehending ability, although the evaluation words "有天赋" is an absolutely positive evaluation literally. Besides, it ought to be notice the substitution needs to consider collocation relation instead of substituting with antonym simply. For example, "美味" is one of antonyms for "恶心", where "美味" means "*delicious*" and "恶心" means "*disgusting*", but "美味" is not supposed to collocate with "男人", which means "*man*", referring to the context in fourth sentence, and it is supposed to be substituted with "优雅" instead, which means "*elegant*", as shown in fifth sentence. Furthermore, the whole sentence needs to be semantically smooth while to expand it with an ironic evaluation when necessary, as shown in sixth sentence.

In this paper, we present a textual adversarial attacking method orienting the NLP task of sentiment analysis by rewriting a straightforward sentence into an ironic sentence, namely IAE (Irony-based Adversarial Examples). To the best of our knowledge, we are the first to use irony for textual adversarial examples generation. We summarize our major contributions as follows:

- We propose IAE, a strategy based on the concept of a rhetorical device called irony for generating textual adversarial examples, which does not need to prepare irony corpus.
- We show that the performance of various deep learning models substantially drops for sentiment analysis tasks when attacked by IAE.
- We show that humans are only mildly or not at all affected by irony in contrast to NLP systems.

II. LITERATURE REVIEW

Our work connects to two strands of literature: textual adversarial examples and irony generation.

A. TEXTUAL ADVERSARIAL EXAMPLES

Existing textual adversarial attack models can be categorized into character-level, word-level, and sentence-level according to the perturbation levels of their adversarial examples.

Character-level attacks disrupt the process of converting natural language text into numerical representations that computers can process, thereby causing model decision shifts. The manifestation of character-level attacks varies across different linguistic environments. In English, character-level attacks often exploit visual perturbations, such as inserting [12], deleting, swapping, and modifying [8] letters within words to create artificially constructed spelling errors. In the Chinese context, handwriting errors on paper do not occur in electronic input based on input methods. Therefore, character-level attacks in the Chinese environment often manifest as the use of homophones for substitution [13], [14] or visual decomposition of characters [15].

Word-level adversarial attacks achieve a shift in the semantic vector of the sample by perturbing the input sample at the word level, causing it to cross the decision boundary and thus leading to incorrect model outputs. Word substitution, as the core method of this strategy, includes various word replacement means such as word vector similarity [6], synonyms [9], and language model scoring [16]. Word-level adversarial attacks do not break the grammatical rules of the text and retain the original semantics to the greatest extent, thus performing better in terms of adversarial text quality and attack success rate. Coupled with the use of language models for control, it also ensures the fluency and smoothness of adversarial texts. Among them, text attacks based on synonym substitution have strong semantic retention and grammatical coherence, belonging to the most threatening category of text adversarial attacks, which have attracted widespread attention from researchers.

Sentence-level adversarial attacks treat the entire original input sentence as the object of perturbation, carefully reconstructing the text content, that is, generating adversarial text that has the same semantics as the original input but causes the victim model to make decision errors. Common sentence-level adversarial attack methods include encoding and then re-decoding [17], adding irrelevant sentences [18], paraphrasing [19], etc.

B. IRONY GENERATION

The field of irony generation, particularly within the Chinese linguistic context, remains largely unexplored, with limited research and development dedicated to this area. Zhu et al. [20] proposed a novel method that integrates reinforcement learning with style transfer techniques to generate ironic text. Their approach relies on a carefully designed reward system to guide the model towards producing text that effectively conveys irony. This method demonstrates the potential of combining advanced machine learning techniques with stylistic adjustments to achieve the nuanced expression of irony. Veale [21] took a different route by exploring knowledge-based systems and shallow linguistic techniques, which they term ''mere re-generation,'' for irony generation. This approach leverages existing knowledge structures and simple linguistic manipulations to introduce ironic elements into the text. While this method may not delve deeply into the complexities of language, it offers a more straightforward and potentially more accessible avenue for irony generation. In the closely related domain of sarcasm, Mishra et al. [22] presented a framework that utilizes reinforced neural sequence-to-sequence learning coupled with information retrieval strategies for sarcasm generation.

To the best of our knowledge, our work represents the first instance of leveraging irony for the generation of textual adversarial examples. This application of irony in adversarial machine learning is groundbreaking, as it introduces a new dimension to the field of natural language processing security. It serves as a testament to the importance of understanding and incorporating advanced linguistic features, such as irony, into machine learning models to enhance their resilience against adversarial attacks.

III. PROBLEM STATEMENT

We assume access to a corpus of labeled sentences $D = \{(s_1, p_1), \ldots, (s_n, p_n)\}$, where s_i is a sentence and $p_i \in L$, the set of possible emotional polarity, i.e., $L = \{\text{positive}, \text{negative}\}$. We define $s^p = (c, e, d)$, a sentence with emotional polarity p, where c is the central word of the sentence, e is the evaluation word that evaluating the central word c, and d is the detailed description of the evaluation. On this basis, we define emotional sentence s^p as a straightforward sentence or an ironic sentence while the evaluation e have emotional polarity p', while collocating with c, and $p = p'$, or $p \neq p'$.

Generally, the irony is a negative sentence exhibiting positive evaluation. Thus, our goal is to build a model that takes as input sentence s, a negative emotional sentence exhibiting negative evaluation e^{neg}, and outputs a sentence s' that retains the negative emotional polarity while exhibiting positive evaluation e^{pos}. Note that the concept of evaluation word we use is not equivalent to the sentiment word while sentiment word is an adjective with a clear emotional polarity. The emotional polarity of an evaluation word should be determined by the central word with which the evaluation word collocates.

IV. APPROACH

In this section, we detail our irony-based textual adversarial attacking method, incorporating three parts: 1) an extractor of collocations between nouns and adjectives, 2) a strategy for evaluation word substitution, and 3) a strategy for ironic evaluation sentence generation. An overview of our IAE generator is shown in Fig. 1. Generally, it takes straightforward text as inputs and outputs ironic text. First, the central word and relevant evaluation word will be located, and then the evaluation word will be substituted with an opposite evaluation word among all possible alternatives, Finally, an appropriate ironic evaluation sentence, determined by local model, will be appended to the text for strengthening the effect of irony.

Next, we describe the details of each component of IAE generator.

A. COLLOCATION EXTRACTOR

We design a collocations extractor to establish noun-adjective collocations tables, which also reveals probable emotional polarity between a noun with all collocated adjectives, as a library of alternatives for evaluation word substitution (see section IV-B).

A host of observations were made on Chinese corpus with part-of-speech tagging and dependency parsing, and we found the noun-adjective collocations in a Chinese sentence are supposed to form the following two kinds of dependencies: 1) a subject-verb structure, or 2) an attributive structure (see examples in Table 2). Note that the results of dependency parsing in Chinese may be different from English due to the differences in the two kinds of syntax rules. e.g., the words ''*weather*'' and ''*good*'' are supposed to form a subject-predicative in English instead of a subject-verb.

Then we can extract plenty of collocations from a large corpus through the observations above, but the next key question is how to determine the emotional polarity of each noun-adjective collocation. Although we can use advanced sentiment analysis models to determine the overall emotional polarity of the sentence from which a noun-adjective collocation extracted, it is no guarantee the emotional polarity of a noun-adjective collocation will be consistent with the

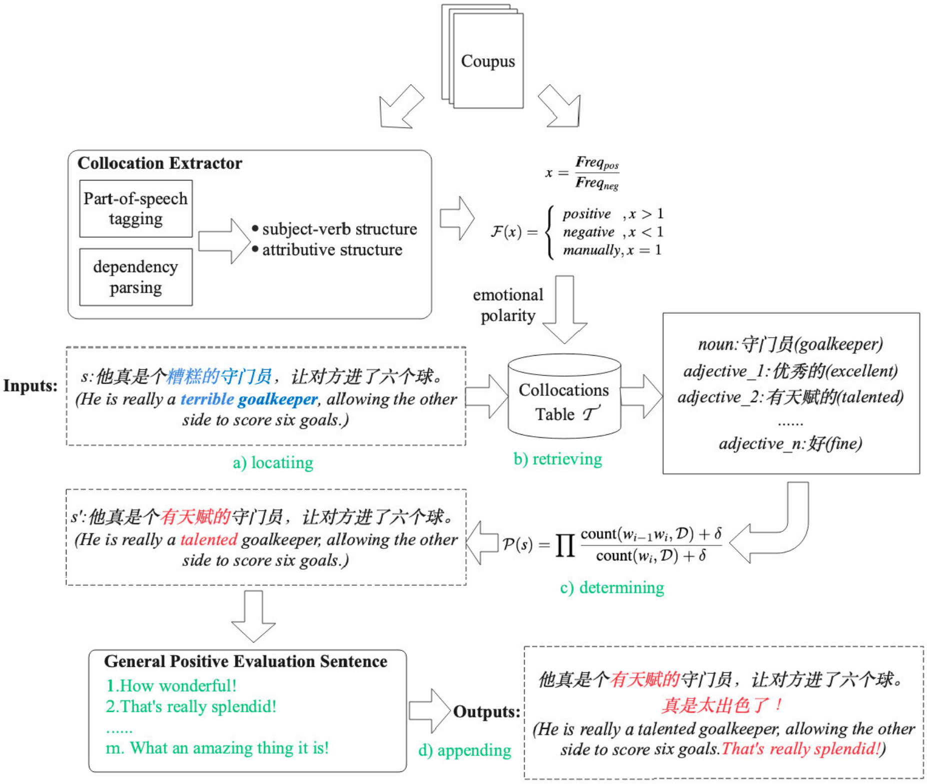
Coupus
Collocation Extractor
Part-of-speech tagging
dependency parsing
• subject-verb structure
• attributive structure
$x = \frac{Freq_{pos}}{Freq_{neg}}$
$\mathcal{F}(x) = \begin{cases} positive & ,x > 1 \\ negative & ,x < 1 \\ manually, & x = 1 \end{cases}$
emotional polarity
Inputs:
s:他真是个糟糕的守门员，让对方进了六个球。
(He is really a terrible goalkeeper, allowing the other side to score six goals.)
a) locatiing
Collocations Table $\mathcal{T}$
b) retrieving
noun:守门员(goalkeeper)
adjective_1:优秀的(excellent)
adjective_2:有天赋的(talented)
......
adjective_n:好(fine)
s′:他真是个有天赋的守门员，让对方进了六个球。
(He is really a talented goalkeeper, allowing the other side to score six goals.)
$\mathcal{P}(s) = \prod \frac{\text{count}(w_{i-1}w_i, \mathcal{D}) + \delta}{\text{count}(w_i, \mathcal{D}) + \delta}$
c) determining
General Positive Evaluation Sentence
1.How wonderful!
2.That's really splendid!
......
m. What an amazing thing it is!
Outputs:
d) appending
他真是个有天赋的守门员，让对方进了六个球。
真是太出色了！
(He is really a talented goalkeeper, allowing the other side to score six goals.That's really splendid!)

FIGURE 1. An overview of our proposed IAE generator.

whole sentence. But, intuitively, the emotional polarity of a collocation should probably be positive if it mostly appears in sentences with a positive overall emotional polarity rather than negative. Hence, the polarity of collocation can be inferred by the following formulas:

$$x = \frac{\mathrm{Freq}_{\mathrm{pos}}}{\mathrm{Freq}_{\mathrm{neg}}} \tag{1}$$

$$F(x) = \begin{cases} \text{positive}, & x > 1 \\ \text{negative}, & x < 1 \\ \text{manually}, & x = 1 \end{cases} \tag{2}$$

where $\mathrm{Freq}_{\mathrm{pos}}$ and $\mathrm{Freq}_{\mathrm{neg}}$ are the frequencies of a collocation appearing in sentences with an emotional polarity of positive or negative respectively. The emotional polarity of a collocation is supposed to be positive when $x > 1$, or negative when $x < 1$, or decided manually when the result of x happens to be 1.

Therefore, the noun-adjective collocations table, denoted as T, can be established by collecting collocations by dependency parsing and inferring their emotional polarities by counting and comparing the numbers of each emotional polarity of the sentences in which they occur.

B. EVALUATION WORD SUBSTITUTION

The strategy for evaluation word substitution is the most important procedure to convert a straightforward sentence s to an ironic sentence s' while s' has the evaluation word e with emotional polarity p', which is opposite to the emotional polarity p of the whole sentence. Next, we describe our evaluation word substitution step by step.

1) LOCATING

At the very beginning, the pairs of central word and relevant evaluation word are located by using part-of-speech tagging and dependency parsing together, which is similar to the strategy of extracting noun-adjective collocation (see secttion IV-A).

2) RETRIEVING

The alternatives are retrieved among the table T by using central word c as an index. The whole procedure will terminate and return a general evaluation word (e.g., ''不错'', which is analog to ''fine'' in English) as the result while the central word does not exist or none of the positive evaluation words are retrieved.

3) DETERMINING

To determine what alternative evaluation word to substitute original, our strategy is to evaluate the quality (i.e., probability of sentence) of all alternative sentences S' by N-gram language model while combining any possible collocation of central word and alternative evaluation word. Formally, for any $s' \in S'$, the probability is calculated

TABLE 2. Examples of sentences containing noun-adjective collocations and dependencies.

Sentences	Collocations	Dependencies
天气这么好，应该出去透透空气。 The weather is so good for enjoying fresh air.	天气，好 weather, good	subject-verb
那个男人真帅。 That man is so handsome.	男人，帅 man, handsome	subject-verb
优美的音乐可以给人们带来享受。 Beautiful music can bring people enjoyment.	音乐，优美 music, beautiful	attributive
她招待我们吃了一顿可口的午餐。 She served us a delicious lunch.	午餐，可口 lunch, delicious	attributive

by the following formula:

$$P(s) = \prod \frac{\text{count}(w_{i-1}w_i, D) + \delta}{\text{count}(w_i, D) + \delta} \tag{3}$$

where w_i is the i-th word in s', $w_{i-1}w_i$ is the sequence composed of w_{i-1} and w_i sequentially, count$(., D)$ denotes the numbers of times a word or a sequence appears in D, and δ is an additive smoothing parameter for the situation that some words are just not appearing in D. In practice, the smoothing parameter δ can be set to 1 empirically. The alternative sentence s' with the highest probability among S' will be determined as the result of evaluation word substitution.

C. IRONIC EVALUATION APPENDING

Reversing the result of sentiment analysis by substituting the evaluation alone is often difficult while the context still exhibits original emotional polarity. But this problem can be solved by appending an evaluation, which is opposite to the polarity of real emotion for strengthening the ironic effect.

It is easy to construct positive evaluations by composing positive adjectives and other grammatical constituents according to sentence patterns. However, the problems are how to choose an evaluation and how to guarantee the semantic smoothness of the whole sentence after evaluation appending.

Our strategy is to construct general positive evaluations, which can collocate with almost objects and guarantee the semantic smoothness, as much as possible, and then to determine an evaluation appending to s'.

Inspired by the substitute black box attack (SBA) [23] which is utilizing the transferability of adversarial examples, we consider training the local model to substitute the victim model, then testing each alternative on the local model, and finally selecting the evaluation while the local model outputs a wrong prediction after appending. For the case that there is no effective adversarial example on the local model, we consider choosing the longest one.

After determining the ironic evaluation, which is supposed to append to the sentence s', the final IAE is generated completely.

V. EXPERIMENTS AND RESULTS

In this section, we conduct comprehensive experiments to evaluate our IAE on the tasks of sentiment analysis.

A. DATASETS AND VICTIM MODELS

We evaluate our IAE on the public reviews of Meituan[1] and Amazon. The five-star and one-star reviews are taken as positive and negative text respectively. Because our IAE is only applicable to the examples with negative emotional polarity, we randomly select 500 examples with negative emotional polarity from each dataset as the test set, and then we divide the remaining examples into two balanced parts for training local and victim models. Details of the datasets are shown in Table 3, where "Class #" refers to the number of labels, "Max. #W" means maximum length of sentences (number of words), "Min. #W" means minimum length of sentences (number of words), and "Avg. #W" means average length of sentences (number of words), "P. #" and "N. #" signify the number of text exhibiting positive and negative emotional polarity respectively.

Besides, for comprehensive noun-adjective collocations extracting (see section IV-A), we collected 30111 nouns and 114383 related collocations from serveral Chinese corpus, including reviews on Meituan and Amazon, Sina weibo[2] comments, and online News corpus. For each noun, there are 1115 collocations at most and 1 collocation at least, with an average of 3.7.

We choose three popular models for text classification, namely TextCNN [24], Bidirectional LSTM (BiLSTM) [25] and a fine-tuned BERT [26], used for evaluating our IAE. TextCNN has three convolutional filters of different kernel sizes (3, 4, 5), and their outputs are concatenated, pooled and fed to a fully-connected layer followed by an output layer. BiLSTM is composed of a 128-dimenional bidirectional LSTM layer, a dropout layer using a drop rate of 0.5, and an output layer. BERT is obtained by fine-tuning the Chinese BERT-Base model with 12-layer, 768-hidden, and 12-heads released by Google. The optimizer, learning rate, and loss function of all models are set to adam, 0.01, and cross-entropy respectively. Besides, we implement Chinese word segmentation, part of speech tagging, and dependency parsing using the third-party library released by Harbin Institute of technology [27].

[1] Meituan is a platform for ordering takeaway, which contains positive and negative user reviews.

[2] A twitter like Chinese online platform.

TABLE 3. Statistics for the datasets.

Dataset	Class #	Max. #W	Min. #W	Avg. #W	P. #	N. #
Meituan	2	237	2	18.96	6000	6500
Amazon	2	858	2	23.71	6000	6500

TABLE 4. Performance of victim models under attacking of IAE and two baseline methods on Meituan review dataset.

Dataset	Method	Local Model	Victim model			WMD
			TextCNN	BidLSTM	Bert	
Meituan	Origin	N/A	0.885	0.894	0.934	N/A
	Visual-based	TextCNN	0.312	0.702	0.880	1.497
		BidLSTM	**0.260**	0.670	0.892	1.787
		Bert	0.610	0.818	0.860	0.384
	Homonym-based	TextCNN	0.326	0.694	0.854	1.551
		BidLSTM	0.304	0.698	0.856	1.784
		Bert	0.606	0.832	0.846	0.411
	Ours	TextCNN	0.464	**0.204**	0.456	1.197
		BidLSTM	0.324	0.416	0.542	1.041
		Bert	0.700	0.876	**0.370**	0.808
Amazon	Origin	N/A	0.900	0.936	0.944	N/A
	Visual-based	TextCNN	0.564	0.664	0.870	2.570
		BidLSTM	0.346	0.400	0.914	2.643
		Bert	0.860	0.830	0.836	2.480
	Homonym-based	TextCNN	0.648	0.702	0.844	2.559
		BidLSTM	0.544	0.550	0.870	2.605
		Bert	0.842	0.826	0.834	2.472
	Ours	TextCNN	**0.338**	**0.322**	0.602	2.344
		BidLSTM	0.720	0.728	**0.538**	2.341
		Bert	0.880	0.886	0.670	2.379

TABLE 5. Human evaluation of emotional correctness and grammar smoothness.

Dataset	Emotional Correctness		Grammar Smoothness	
	Origin	IAE	Origin	IAE
Meituan	91	86	4.25	3.50
Amazon	90	84	4.50	3.75

B. BASELINE METHODS

We implement two baseline methods based on important word substitution and compared them with ours for proving the contribution of this work. The two baseline methods are 1) visual-based substitution [8], which means substitute important words with visual similar chart, and 2) homonym-based substitution [13], which means substitute important words with others pronounced the same way but have different meanings. The important words refer to the words in the input text that make the most contribution to the model decision and the calculation algorithm of important words adopts [28].

C. ATTACK PERFORMANCE

The attack performance results of our IAE and two baseline methods are shown in Table 4. Note that only examples labeled with negative are used for test as the adversarial attack based on irony is only applicable to the examples with negative emotional polarity. We observe the adversarial examples generated by our irony-based attack cause the victim models to fail more seriously than the baseline methods in most conditions in most conditions. Specifically, the visual-based and homonym-based attack can hardly fool Bert models while our method can cause the accuracy of Bert from 89.8% to 37.0% at most, besides, the Word Mover's Distances [29] between our IAE and clean examples are always smaller than those between adversarial examples generated by baseline methods and clean examples.

D. HUMAN EVALUATION

We ask 4 students with native Chinese language skill to evaluate the emotional correctness and semantic smoothness of successful IAE generated from Meituan and Amazon reviews. Specifically, we randomly select 100 IAE and 100 clean examples, and every student needs to evaluate the mixture of them.

For evaluating emotional correctness, each student evaluates the true emotional polarity of each example and it is annotated as positive (negative) if two or more students evaluate a example as positive (negative). An extra human evaluator would participate in the evaluation if there are equal numbers of different evaluation on emotional polarity.

For evaluating semantic smoothness, each student scores the semantic smoothness of each example with Likert scale ranging from 1 to 5 while 1 and 5 mean the semantics of a example is completely confused or fluent separately. We summarized the evaluation of all students and averaged the semantic smoothness of the IAE and the clean examples respectively.

The results are shown in Table 5 and it shows that a lightly lower emotional correctness and semantic smoothness in IAE than clean examples, but the emotional correctness and semantic smoothness of IAE still reach 86 and 3.75 respectively.

VI. DISCUSSION

We studied how to regard irony as a textual adversarial perturbation in Chinese and it proved effective in sentiment analysis. There are differences between Chinese and other languages in grammar and habits, however, irony, as a rhetorical device in almost all languages, could be utilized as a general way of textual adversarial perturbation.

The experiment of training the local model for generating effective adversarial examples also reveals some properties of transferability. First, the transfers between two models are non-symmetric. As we can see, the accuracy of victim model BERT is 54.2% when generated IAE from local model BidLSTM, however, the accuracy of victim model BidLSTM is 87.6% when generated IAE from local model BERT while testing on Meituan reviews dataset. It is similar to the findings in the study of the transferability of image adversarial examples [30], even though we focus on the text field. Second, the adversarial examples generated from the high-accuracy models may be less transferable. As we can see, BERT is the most accurate model among all models we use, however, the adversarial examples generated from BERT hardly mislead other models.

We also found there are three major types of weaknesses in our methods, which affect the attacking performances. For analysis of the weaknesses, we sampled 100 failed IAE which mislead the victim model unsuccessfully or lose original sentiment. We found that 26% of the failures are due to the long length of input text which is more than 50 Chinese characters, 38% of the failures are due to the weak correlation between evaluative sentence and context, 29% of the failures are due to the imperfection of part-of-speech tagging and dependency parsing tools, and the remaining 7% of the failures have no significant type.

The first type of failure is due to the obvious fact that the longer the text, the more negative content it contains, so it is difficult to change the label of model prediction by substituting an evaluation word or appending a generally positive evaluation sentence.

The second type of failure is due to the weak correlation between the evaluative sentence and the context description. For example, for the sentence "菜真的很难吃, 还是去其他店吃好些" (The food is really unpalatable, and it's better to go to another restaurant), where the context is not a correlational detail description to the evaluation of food, it is inappropriate to substitute the negative evaluation word "难吃" (unpalatable) to a positive word "好吃" (delicious) otherwise the emotional polarity of the text will change completely.

The third type of failure is due to the dependency analysis tools, which is unable to analyze the dependency correctly all the time, while it is necessary to locate the evaluation word with part-of-speech tagging and dependency parsing.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have introduced Irony-based Adversarial Examples (IAE), a novel method for generating adversarial text by transforming straightforward sentences into ironic ones. Our research has made several significant contributions to the field of adversarial attack. Firstly, we have introduced IAE as a strategy for generating textual adversarial examples that leverages irony. This method is innovative in that it does not depend on pre-existing irony corpora, thereby offering a flexible instrument for creating adversarial text across a spectrum of NLP tasks. Secondly, we have demonstrated empirically that the performance of several deep learning models on sentiment analysis tasks is markedly compromised when confronted with IAE attacks. This result highlights the vulnerability of current NLP systems to adversarial manipulations facilitated through irony. Thirdly, we have compared the effects of IAE on human judgment versus NLP systems, revealing a notable difference in susceptibility. Our findings indicate that humans are relatively more resilient to the influence of irony in text, contrasting with the performance of NLP models.

Our future work will focus on enhancing the performance of IAE in longer texts and improving its generalization capabilities across different languages. This will involve addressing the complexities associated with maintaining ironic integrity over extended passages and adapting to the nuances of various linguistic contexts. Additionally, we are intrigued by the prospect of integrating more rhetorical devices into textual adversarial perturbations, beyond irony. For instance, exploring the use of metaphors to disrupt machine reading comprehension presents an exciting avenue for further research. By expanding the repertoire of rhetorical strategies employed in adversarial text generation, we aim to deepen our understanding of the interplay between language, context, and machine learning models, ultimately contributing to the development of more robust and nuanced NLP systems.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–11.

[2] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.

[3] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 2021–2031, doi: 10.18653/v1/d17-1215.

[4] Q. Wang, B. Zheng, Q. Li, C. Shen, and Z. Ba, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 896–908, 2021, doi: 10.1109/TIFS.2020.3026543.

[5] Y. Belinkov and Y. Bisk, "Synthetic and natural noise both break neural machine translation," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=BJ8vJebC-

[6] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, ''Is BERT really robust? A strong baseline for natural language attack on text classification and entailment,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 8018–8025. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6311

[7] 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 (Short Papers)*, vol. 2, Melbourne, VIC, Australia, 2018, pp. 31–36.

[8] S. Eger, G. G. Sahin, A. Rücklé, J.-U. Lee, C. Schulz, M. Mesgar, K. Swarnkar, E. Simpson, and I. Gurevych, ''Text processing like humans do: Visually attacking and shielding,'' in *Proc. Conf. North*, Minneapolis, MN, USA, 2019, pp. 1634–1647, doi: 10.18653/v1/n19-1165.

[9] 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, 2019, pp. 1085–1097, doi: 10.18653/v1/p19-1103.

[10] 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. (Long Papers)*, vol. 1, 2018, pp. 1875–1885, doi: 10.18653/v1/n18-1170.

[11] D. Pruthi, B. Dhingra, and Z. C. Lipton, ''Combating adversarial misspellings with robust word recognition,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5582–5591, doi: 10.18653/v1/p19-1561.

[12] B. Formento, C. S. Foo, L. A. Tuan, and S. K. Ng, ''Using punctuation as an adversarial attack on deep learning-based NLP systems: An empirical study,'' in *Proc. Findings Assoc. Comput. Linguistics: EACL*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 1–34.

[13] W. Wang, R. Wang, L. Wang, and B. Tang, ''Adversarial examples generation approach for tendency classification on Chinese texts,'' *Ruan Jian Xue Bao/J. Softw.*, vol. 30, pp. 2415–2427, Apr. 2019, doi: 10.13328/j.cnki.jos.005765.

[14] C. Nuo, G.-Q. Chang, H. Gao, G. Pei, and Y. Zhang, ''WordChange: Adversarial examples generation approach for Chinese text classification,'' *IEEE Access*, vol. 8, pp. 79561–79572, 2020, doi: 10.1109/ACCESS.2020.2988786.

[15] H. Ou, L. Yu, S. Tian, and X. Chen, ''Chinese adversarial examples generation approach with multi-strategy based on semantic,'' *Knowl. Inf. Syst.*, vol. 64, no. 4, pp. 1101–1119, Apr. 2022.

[16] H. Zhang, H. Zhou, N. Miao, and L. Li, ''Generating fluent adversarial examples for natural languages,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5564–5569, doi: 10.18653/v1/p19-1559.

[17] W. Han, L. Zhang, Y. Jiang, and K. Tu, ''Adversarial attack and defense of structured prediction models,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2327–2338.

[18] 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.* Stockholm, Sweden: IJCAI, Jul. 2018, pp. 4208–4215, doi: 10.24963/ijcai.2018/585.

[19] Y. Xu, X. Zhong, A. Jimeno Yepes, and J. H. Lau, ''Grey-box adversarial attack and defence for sentiment classification,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 1–10.

[20] M. Zhu, Z. Yu, and X. Wan, ''A neural approach to irony generation,'' 2019, *arXiv:1909.06200*.

[21] T. Veale, ''A massive sarcastic robot: What a great idea! two approaches to the computational generation of irony,'' in *Proc. 9th Int. Conf. Comput. Creativity (ICCC)*F. Pachet, A. Jordanous, and C. León, Eds., Salamanca, Spain: Association for Computational Creativity, Jun. 2018, pp. 120–127.

[22] A. Mishra, T. Tater, and K. Sankaranarayanan, ''A modular architecture for unsupervised sarcasm generation,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 6143–6153, doi: 10.18653/v1/d19-1636.

[23] 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.*, Abu Dhabi, United Arab Emirates, Apr. 2017, pp. 506–519, doi: 10.1145/3052973.3053009.

[24] Y. Kim, ''Convolutional neural networks for sentence classification,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751, doi: 10.3115/v1/d14-1181.

[25] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, ''Supervised learning of universal sentence representations from natural language inference data,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 670–680, doi: 10.18653/v1/d17-1070.

[26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[27] W. Che, Y. Feng, L. Qin, and T. Liu, ''N-LTP: An open-source neural language technology platform for Chinese,'' 2020, *arXiv:2009.11616*.

[28] 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.*, San Diego, CA, USA, 2019, pp. 1–15. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/textbugger-generating-adversarial-text-against-real-world-applications/

[29] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, ''From word embeddings to document distances,'' in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966. [Online]. Available: http://proceedings.mlr.press/v37/kusnerb15.html

[30] L. Wu and Z. Zhu, ''Towards understanding and improving the transferability of adversarial examples in deep neural networks,'' in *Proc. 12th Asian Conf. Mach. Learn. (ACML)*, vol. 129, S. J. Pan and M. Sugiyama, Eds., Bangkok, Thailand, 2020, pp. 837–850. [Online]. Available: http://proceedings.mlr.press/v129/wu20a.html

XIAOYIN YI received the M.S. degree in computer science from Chongqing University of Posts and Telecommunications, China, where she is currently pursuing the Ph.D. degree. She is currently a Teacher with Chongqing College of Mobile Communication. Her research interests include cybersecurity within AI and AI security.

JIACHENG HUANG received the M.S. degree in instructional technology from Hubei Normal University, China. He is currently pursuing the Ph.D. degree with Chongqing University of Posts and Telecommunications, China. His research interests include cybersecurity within AI, natural language processing, and AI security.