Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research on the Generation of Tibetan Word-levelAdversarial Samples Based on Morphosyntactic Linkage

Btsan lhaMgon
Xizang University
Qun Nuo
q_nuo@utibet.edu.cn
Xizang University
Nyima Tashi
Xizang University

Research Article

Keywords: Syntactic linkage, function words, word-level granularity, adversarial examples

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10186980/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Research on the Generation of Tibetan Word-level Adversarial Samples Based on Morphosyntactic Linkage

Btsan lhaMgon[1,2,3], Qun Nuo[1,2,3*], Nyima Tashi[1,2]

[1], School of Information Science and Technology, Xizang University , Lhasa, 850000, Xizang Autonomous Region, China.
[2], The State Key Laboratory of Tibetan Intelligence, Lhasa, 850000, Xizang Autonomous Region, China.
[3], Collaborative Innovation Center for Xizang Informatization by MOE and Xizang Autonomous Region , Lhasa, 850000, Xizang Autonomous Region, China.

*Corresponding author(s). E-mail(s): q_nuo@utibet.edu.cn;
Contributing authors: zlg@stu.utibet.edu.cn; q_nuo@utibet.edu.cn;

Abstract

Tibetan word substitution is a common technical strategy in text adversarial defense and data augmentation. This paper addresses the issue of neglecting Tibetan grammatical norms for function words during the core keyword substitution process in the field of Tibetan word-level text adversarial attacks and defense, proposing a novel text adversarial sample generation method TWLASBMSI that conforms to Tibetan grammatical rules through morphosyntactic linkage. The method consists of four core steps: First, using a Tibetan sememe knowledge base, the content in the original Tibetan text is randomly matched and replaced once with words of higher semantic similarity; second, employing a function word attachment detection and correction algorithm based on Tibetan grammatical categories to ensure automatic and consistent adaptation and updating of function words such as genitive, ergative, locative, and personal noun suffixes in the replaced text, thereby guaranteeing the generation effectiveness and text quality of Tibetan adversarial samples. Finally, the generated Tibetan adversarial samples are input into the target model to carry out attacks. Experimental results show that the TWLASBMSL attack method leads to an average accuracy drop (ADV) of 2.69%, an average attack success rate (ASR)

of 28.83%, and an average Levenshtein distance (LD) of 6.80% on Tibetan sentiment analysis and text classification tasks using TiBERT, CINO-base-v2, and CINO-large target models. In the task of automatic adaptation and error correction for Tibetan function words, the function word error detection rate, overall accuracy, and complete error sentence correction success rate all reached 100%. Overall, the experimental results indicate that the proposed morphosyntactically linked text adversarial sample generation method TWLASBMSL not only effectively misleads models into making incorrect classification decisions but also ensures the grammatical logical rigor of Tibetan adversarial samples.

Keywords: Syntactic linkage, function words, word-level granularity, adversarial examples

1 Introduction

After more than 40 years of research, Tibetan intelligent information processing has achieved corresponding academic research results in machine translation [1], computer vision[2], speech processing [3], Tibetan medicine [4]. In recent years, due to the gradual improvement of computing power and data, Tibetan large models such as DEEPZANG, YangGuang QingYan and ZHIDA AI have become increasingly common, achieving significant technological progress. However, the security and robustness of Tibetan large models cannot be ignored. For example, large models learn human language knowledge through massive amounts of data, and then fine-tune them with human cultural values to form a language model that can understand and generate high-quality content. Due to the learning mechanism of the model and data bias, black box characteristics, etc., large models have defects such as illusion [5] and easy generation of harmful content, and are susceptible to adversarial attacks [6]. Text adversarial and defense methods are one of the key means to solve such problems [7]. Adversarial attacks and defenses are achieved by carefully perturbing the text with adversarial samples, aiming to improve the robustness and security of large models under small perturbations and ensure the safe implementation of the model.

In order to enhance the defense capability of Tibetan language large model in adversarial attack scenarios and improve robustness and security, text adversarial and defense is a mature and commonly used method at present [8]. Keyword replacement [9] is the most commonly used and most basic attack and defense method. Existing Tibetan keyword replacement methods include syllable replacement and character-level replacement. The first method is to select the Tibetan syllable with the highest score after judging the salience of the syllable and replacing it with the syllable with the highest semantic vector. The second method is to select the Tibetan syllable with the highest score after judging the salience of the Tibetan syllable and generate Tibetan adversarial samples by inserting, deleting, replacing, and other multi-granular perturbation methods. The two attack methods mentioned above provide a feasible research method for Tibetan text adversarial, but in actual adversarial samples, there are problems such as imprecise syntactic logic and the need to improve attack accuracy.

Specifically, regardless of whether the attack granularity is at the character, syllable, or word level, if the replacement strategy only achieves superficial perturbations at the character, syllable, and word levels without constraining syntactic legality, the adversarial samples generated after minor formal modifications will have quality issues such as invalid grammatical rules and broken sentence logic. For example, in the Tibetan sentence " བཀྲ་ཤིས་ཀྱིས་དཔེ་ཚ་སྒོག་གིན་ཡོད། " (Tashi is reading), this sentence replaces the syllable with the highest semantic similarity based on the salience importance of the syllable, replacing the syllable " སྒོག "(read) Replacing " འདོན་ "(read), which has the highest semantic similarity, yields the generated adversarial sample " བཀྲ་ཤིས་ཀྱིས་དཔེ་ཚ་འདོན་གིན་ཡོད། " (Tashi is reading). If we consider the relevance of adversarial strategies between Chinese and English... The attack method and the work of generating adversarial samples have been completed. However, in the Tibetan adversarial strategy, the grammatical logic of the sample sentences after replacing characters, syllables, and words has deviated. When the syllable " སྒོག " (read) is replaced with the syllable " འདོན " (read), which has the highest semantic similarity, according to the Tibetan text rules, the function word after the syllable " འདོན " (read) is also changed to " འདོན " (read) according to the function word concatenation rules, resulting in the complete sentence " བཀྲ་ཤིས་ཀྱིས་དཔེ་ཚ་འདོན་ཀྱིན་ཡོད། " (Tashi is reading). Therefore, this paper proposes a Tibetan Word-Level Adversarial Sample Generation Based on Morphological Interaction (TWLASBMSI) method. This method leverages the Tibetan Sememe Knowledge Baseby randomly matching content from the original Tibetan text and performing single-word replacements with semantically similar terms. It also ensures consistency in the replacement of Tibetan genitive, ergative, locative, and personal noun suffix particles, thereby guaranteeing the effectiveness and text quality of the generated Tibetan adversarial examples.

The main contributions of this paper are reflected in the following points:

- Proposed a word-level adversarial example generation method based on morphosyntactic linkage, effectively enhancing the robustness of Tibetan language models under white-box perturbation attacks.
- Established criteria or algorithms for the automatic adaptation and updating of Tibetan function words, including perturbation replacement of Tibetan characters, syllables and words.
- Constructed a Tibetan Sememe Knowledge Basewith a total scale of 32,745 entries, and used it to conduct experiments on adversarial attacks at the word level in Tibetan text.

2 Related work

In 2014, Szegedy et al [10] first applied the adversarial attack method to the study of image classification. After making tiny perturbations to the pixels of the image that were invisible to the naked eye, the model made incorrect classification results. Subsequently, this method was applied to text adversarial attacks starting in 2017. At that time, Jia et al [11] added a suitable amount of interference sample sentences

to the SQuAD reading comprehension questions, causing the SQuAD reading comprehension system to output incorrect results. Jia et al [12] pioneered the field of text adversarial attacks. After that, perturbation attack methods of various granular levels such as character level, word level, phrase or clause level, sentence level, and semantic level appeared one after another. For example, in 2017, Ebrahimi et al [13] proposed a gradient-based character perturbation white-box attack method, which achieves text adversarial attacks by modifying, subtracting, deleting, and adding characters; in 2018, Gao et al [14] first proposed a character-level perturbation-based black-box adversarial method, which became the pioneering work of fine-grained text adversarial perturbation research; Jin et al [15] proposed a word-level perturbation-based TextFooler black-box adversarial method, which achieves efficient attacks on models such as BERT; Wang et al [16] proposed a sentence-level controllable adversarial method based on CAT-Gen, which ensures adversarial text with strong semantic consistency; Zheng et al [17] first proposed a dependency syntax adversarial attack method in black-box and white-box scenarios to improve the robustness of the model.

As adversarial attack methods are gradually applied to Chinese text classification and sentiment analysis scenarios, adversarial text research for Chinese has also emerged. Various perturbation attack methods at multiple granular levels, such as character level, word level, and sentence level, have been developed. For example, Zhang Shunxiang et al [18] implemented a character-level black-box attack method by perturbing polyphonic characters and verified the effectiveness of the method on the sentiment classification task; Tong Xin et al [19] proposed a word-level black-box adversarial method, CWord Attacker, which realizes adversarial samples with various perturbation methods such as pinyin and traditional character replacement through the TDS targeted word deletion scoring mechanism; Gong Yan et al [20] constructed an exclusive adversarial search domain based on the two inherent linguistic features of Chinese characters, namely character shape and sound, and proposed a Chinese text classification method based on black-box attack, WordBeguiler; Li Xiangge et al [21] proposed a word-level adversarial attack method, CWordCheater, based on multiple linguistic features of Chinese characters, such as sound, character shape, and punctuation marks. Wu Houyue et al [22] proposed DiffuAdv, an adversarial attack method based on attack-guided diffusion, to address the limitations of existing adversarial methods such as word substitution and word order adjustment.

In recent years, in order to ensure that Tibetan language models can run stably, reliably, and securely in real scenarios, text adversarial attack and defense techniques are also used to improve the robustness and security of the models. However, due to the scarcity of researchers and related results, research on Tibetan text adversarial attacks and defenses is still in its infancy compared to English and Chinese. Currently, the relevant research work carried out in this field only includes perturbation attack methods at the character level and syllable level. For example, Cao et al. [23–25] first proposed a Tibetan black-box adversarial attack method based on syllable-level perturbation, and further proposed TSTricker, a Tibetan syllable and word-level adversarial attack method based on masked language models. Subsequent research relied on the similarity of character shapes in Tibetan syllables to make

the model distinguish confusion and achieve effective text adversarial attacks. Li Jiajun et al [26] proposed DAAttacker, an adversarial sample method at the character level and syllable level, for six perturbation methods such as deletion, insertion, and replacement based on Tibetan syllables.

Most of the Tibetan adversarial attack methods mentioned above are designed based on character-level, word-level, and word-level substitution techniques used in Chinese and English. The attack granularity and methods employed are consistent with current technological trends in the field. However, deviations occur in the Tibetan grammatical logic of the adversarial samples generated after multi-level substitution. Current Tibetan text adversarial and defense methods, besides using multi-level granular substitution to generate adversarial samples, neglect the rules for adding Tibetan function words in the generated adversarial samples. When substituting Tibetan syllables or words in adversarial methods, the subsequent Tibetan genitive, ergative, locative, and person noun suffixes undergo corresponding linkage changes based on the attributes of the replaced language unit. To address these issues, this paper proposes a morphological-syntactic linkage adversarial sample generation method, TWLAS-BMSL, based on the function word addition rules in Tibetan grammar. This method uses a Tibetan Sememe Knowledge Base to select words with higher semantic similarity for random matching and completes a single word substitution. At the same time, the quality of adversarial samples is ensured by using the function word concatenation rules in Tibetan grammar.

3 Tibetan function words and their rules

3.1 The concept of Tibetan function words

In Tibetan grammar, function words usually cannot express lexical meaning. Instead, they must be combined with other linguistic structures in the Tibetan sentence to express the complete semantic relationship between nouns and predicates.

According to statistics, there are 85 function words in Tibetan grammar, of which 20 are free function words and 65 are inflexible function words. When a free function word functions as a sentence component, the structure of the preceding syllable does not restrict its use, and it can be added freely according to the components in the sentence. However, the method of adding inflexible function words is exactly the opposite of that of free function words. When an inflexible function word functions as a sentence component, the structure of the preceding syllable restricts the function word added after it. It must be added in combination with the structure of the preceding syllable and cannot be added arbitrarily.

3.2 Tibetan rules for adding function words

The rules for adding Tibetan function words are the core content of the traditional Tibetan grammar text “The Thirty Verses”. This rule is based on the post-suffix letters of syllables, categorizing function words into free and non-free types. Free function words, such as modal particles “ ནི ”, linking particles “ དང ”, and demonstrative

pronouns “ དེ ”, are not constrained by post-suffix letters and can be added arbitrarily. For example, in the sentence “ ཁ་ཁང་གི་མེ་དེ། ད་དང་ཚོང་ཀྱི་འཛེལ་བ། ” the linking particle“ དང ” and demonstrative pronoun “ དེ ” can be freely used without constraints. Non-free function words, such as genitive, ergative, locative, associative, and pending particles, must follow the variations based on the ten post-suffix letters “ ག་ང་ད་ན་བ་མ་འ་ར་ལ་ས ” and the absence of a post-suffix letter “ འ ”. They require corresponding morphological changes. For instance, the five genitive particles “ ཀྱི་གི་གྱི་འི ” must be selected and determined based on the post-suffix letter of the preceding syllable. In the phrase “ ཅོག་ཅེ་སྟེང་གི་སྐྱུགས་ཀུ ” (the pen on the table), the addition of the function word “ གི ” is determined by the post-suffix letter “ ང ” of the preceding syllable. The rules for adding function words run through the entire process of Tibetan grammar and serve as the foundational principle for ensuring accurate writing, reading, and grammatical expression in Tibetan.

The rules for appending Tibetan function words are determined based on the suffix letters and additional suffix letters in the syllable preceding the function word. Different suffix letters or additional suffix letters in different syllables result in varying rules for appending function words in specific contexts. The specific rules for appending Tibetan function words are summarized in Tables 1 below.

Table 1 Rules for Adding Case Particles in Tibetan

Suffix	Genitive	Agentive	Dative	Suffix for personal nouns	function words
ད	གྱི	གྱིས	དུ	ད	ཞིང/ཞས/ཞེའོ/ཞེ་ན/ཞིག
ན	གྱི	གྱིས	དུ	ད	
བ	གྱི	གྱིས	དུ	ད	
མ	ཡི/ཡི	ཡིས/དེས	-ར/རུ	ད	
འ	གྱི	གྱིས	དུ	ད	
ར	གྱི	གྱིས	དུ	ད	
ལ	གྱི	གྱིས	ཏུ	ད	ཅིང/ཅས/ཅེའོ/ཅེ་ན/ཅིག
ས	གྱི	གྱིས	དུ	ད	
ག	གྱི	གྱིས	ཏུ	ད	
-ད			ཏུ	ད	
ས	གྱི	གྱིས	སུ	ད	ཤིང/ཤས/ཤེའོ/ཤེ་ན/ཤིག
No suffix added	ཡི/ཡི	ཡིས/དེས	-ར/རུ	ད	ཞིང/ཞས/ཞེའོ/ཞེ་ན/ཞིག

Note: In the above “ ད ”, the “-” represents any additional character.

Tables 1 describes the system of adding genitive or agentive,dative, and personal noun suffixes based on each column of characters in the post-suffix column, corresponding item by item, forming a complete correspondence between post-suffixes and function words in phonetic characters.

4 Method

4.1 Problem Definition

This paper focuses on Tibetan word-level perturbation methods to achieve text adversarial attacks. To better utilize perturbed text to attack the target model, this paper, under a white-box attack paradigm, generates new adversarial samples by perturbing the input Tibetan text at the word level.

Let the original Tibetan text to be tested be denoted as "x", and let "F" be the Tibetan target model. By applying word-level perturbation δ to the original Tibetan text "x", an adversarial sample "x_1" is obtained. To complete the random word substitution perturbation for Tibetan sememes, four requirements must be met: First, the perturbation operation on the Tibetan text should generate corresponding adversarial samples through minimal perturbations, making the perturbation imperceptible to humans in terms of both semantic understanding and reading fluency. Second, when generating adversarial samples for Tibetan at the word level, word substitutions must involve words with high semantic similarity. Third, for the replaced Tibetan function words, synchronous and consistent substitutions should be performed, adhering to the collocation rules of function words in Tibetan grammar to ensure the generation effectiveness and text quality of the Tibetan adversarial samples. Fourth, when the adversarial sample "x_1" generated by perturbation is input into the "F" model for testing, the test result should be opposite to that of the original text.

In order to better utilize Tibetan word-level granularity to achieve text adversarial and defensive effects, this paper employs a Tibetan word-level method based on morphosyntactic linkage to implement the attack method TWLASBMSL. The specific attack step architecture is shown in Figure 1.

The Tibetan word-level method for pattern definition first performs random word substitution perturbations based on Tibetan semantic knowledge on the original Tibetan text, then utilizes the Tibetan function word attachment method to reconfigure the adversarial samples of the Tibetan function word structure after substitution, and finally, inputs the generated Tibetan adversarial samples into the target model to carry out the attack.

4.2 Random word perturbation

In the process of generating adversarial samples of Tibetan text at the word level, the use of highly similar semantic primitives to perform precise word-level replacement perturbation on Tibetan sentences can not only preserve the semantics to the greatest extent [27], but also ensure the quality of adversarial samples. This paper takes advantage of the Tibetan Sememe Knowledge Baseand uses the random word perturbation method to replace the content in the original Tibetan text with words with

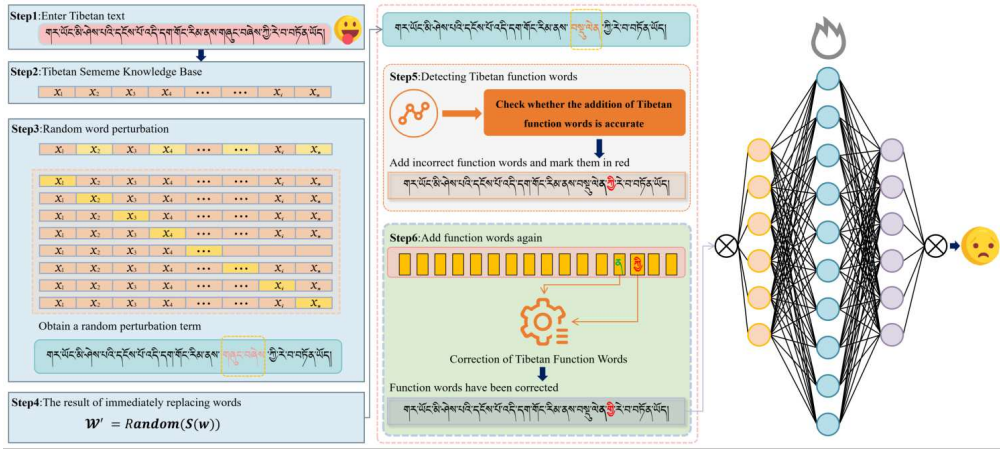


Fig. 1 TWLASBMSL Method Framework

higher semantic similarity through random matching and single replacement, thereby generating high-quality adversarial samples.

The algorithm for random word replacement perturbation based on Tibetan semantic knowledge is shown in Algorithm 1.

4.3 Tibetan function words rearrangement

In order to generate high-quality adversarial examples with few modifications and no grammatical or logical errors, this paper finds Tibetan word-level keywords that have the greatest impact on the prediction of the target model through perturbation localization and replaces them. At the same time, the Tibetan function words following the replaced keywords are dynamically adapted and changed according to the Tibetan function word addition rules, depending on the addition of words after the keyword or whether it is the agent subject.

In Tibetan data augmentation and text adversarial defense, when replacing a word in the text with a candidate word that has high semantic similarity to generate a new text or adversarial sample, we cannot simply copy the data augmentation and text adversarial attack and defense methods of English and Chinese. This is because the deep semantic relationship of Tibetan is organized into sentences through function words [28]. In addition to the differences in sentence structure between Tibetan and Chinese and English, the most significant difference is that there are a large number of function words in Tibetan sentences, which play a crucial role in the components of Tibetan sentences.

When directly using word-level adversarial attacks and defenses in Chinese and English to generate adversarial samples, the replacement operation can be performed as long as the keywords to be replaced in the sentence satisfy the condition of high

Algorithm 1 An algorithm for random word substitution perturbation based on Tibetan semantic knowledge.

Require: A set of Tibetan sentences S, and a set of semantic primitive substitution rules $R = \{(k, \upsilon)\}$, where k is the word to be replaced and ? is the replacement word.

Ensure: The rewritten set of Tibetan sentences S'

Initialize $S' \Leftarrow \emptyset$
Read all lines from rule file, split by delimiter **%%** to construct rule set R
for each sentence $s \in S$ **do**
    $current_s \Leftarrow s$
    $Candidates \Leftarrow \emptyset$ ▷ Match words with surrounding Tibetan separator **?**
    for each rule $(k, \upsilon) \in R$ **do**
        Construct matching pattern: $pattern = (\texttt{?} \leq \texttt{?})\, k\, (\texttt{?} = \texttt{?})$
        if *pattern* matches substring of $current_s$ **then**
            $Candidates \Leftarrow Candidates \cup \{(pattern, \upsilon)\}$
        end if
    end for ▷ Execute random substitution if matched candidates exist
    if $Candidates \neq \emptyset$ **then**
        Randomly select one pair $(p, \upsilon_{\text{set}})$ from *Candidates*
        Replace the first substring matching pattern p in $current_s$ with υ_{set}
    end if
    $S' \Leftarrow S' \cup \{current_s\}$
end for
return S'

semantic similarity, without considering other grammatical logic. However, when generating Tibetan adversarial samples using word-level perturbation, the replacement operation step must not only satisfy the condition of high semantic similarity, but also satisfy the function word concatenation rules in Tibetan grammar. For example, Table 2 shows the changes in grammatical logic when generating word-level adversarial samples in Chinese and English.

Table 2 Examples of word-level adversarial samples in Chinese, English, and Tibetan

English	Blame should	Blame
Chinese	不对/之处/应/多/从/自身/找/原因	不对/之处/应/多/从/自身/找/原因
Tibetan	རྩོད་འཆུག་གི/རྐྱེན/ཆེད/བྱོང/ཀྱི/ར་དང་སྦྱར/ལ/བཅོལ/ན/འགྱོ	མི་འཐུག་པ/ལེ/རྐྱེན/ཆེད/བྱོང/ཀྱི/ར་དང་སྦྱར/ལ/བཅོལ/ན/འགྱོ

In the process of generating word-level adversarial examples for English and Chinese, it is sufficient to replace “mistakes” in English and “ 因由 ” in Chinese with “errors” and “ 缘由 ” respectively, which have the highest semantic similarity, without performing any additional grammatical structure changes. However, the generation of

word-level adversarial examples for Tibetan is exactly different from that for English and Chinese. In the process of generating the examples, not only must the keyword " རོར་འཛུག " be replaced with " མི་འགྱིག་པ " which has a higher semantic similarity, but the usage of function words after the replaced keyword " མི་འགྱིག་པ " must also conform to the grammatical logic of Tibetan.

If the rules for adding Tibetan function words are not accurately followed, the generated adversarial examples may actually affect the original semantics of the text. Therefore, this paper proposes an algorithm for dynamically adapting and linking Tibetan word-level keywords and function words to address the problem of violating the rules for adding Tibetan function words in generated adversarial examples. The specific algorithm process is shown in Algorithm 2.

5 Experiments and Analysis

5.1 Experimental data

The data used in this study include the TNCC Tibetan News Classification dataset released by the Natural Language Processing Laboratory of Fudan University (only the news title classification was used, covering 12 text classification scenarios such as politics, economics, and education), the TU_SA Tibetan Sentiment Analysis dataset released by Professor Qun Nuo's team at Tibet University, the Tibetan Sentiment Analysis dataset published on the Tibetan Natural Language Processing Platform by Qinghai Normal University (hereinafter referred to as "QHNUTSA"), and the Weibo Comment dataset released by Professor Yan Xiaodong's team at Minzu University of China (this dataset covers three sentiment types: positive, neutral, and negative; this study only used positive and negative data for experiments, hereinafter referred to as "MZUCTSA"). These Tibetan datasets exhibit several quality issues: Tibetan texts contain a large number of English characters and special punctuation marks such as ":(" and "ree@.."; the word length of single-line Tibetan sentences is generally short (some consist of only two syllables); there are label mismatches in the data used for Tibetan text classification and sentiment analysis, and the texts contain issues such as incorrect concatenation of two syllables due to the absence of the Tibetan separator " ་ ", resulting in non-standard variant errors (e.g., " འགེན་ ") and misspellings (e.g.," བྱས "). Therefore, to address the aforementioned data issues, this study invited three master's students majoring in Chinese Ethnic Minority Language and Literature to proofread, correct, and standardize the content of the Tibetan datasets item by item.

The overall dataset was stratified and divided into an 80% training set and a 20% test set. The specific details of the experimental data are presented in Table 3 below:

5.2 Target Model

Large-scale models specifically designed for Tibetan are scarce in both industry and academia, and the few existing Tibetan models do not provide external access services. Therefore, this experiment selected TiBERT and CINO pre-trained models that support Tibetan as target models for evaluation on Tibetan sentiment analysis

Algorithm 2 Algorithm for dynamic adaptation and linkage between keywords and function words.

Require: Tibetan original text $X = (w_1, w_2, \ldots, w_i)$, Tibetan function word regularization matching and replacement rule set T, random word replacement perturbation function R based on Tibetan semantic primitives, target model F

Ensure: Generated Tibetan adversarial example X'

function T_RegexReplace(text, T)
  for (pat, rep) $\in T$ **do** ▷ Identify function words in text
text ← RegexReplace(text, pat, rep)
  end forreturn text
end function
function func_word_adj(tokens, i)
  tibetan_B ← tokens[$i+1$].strip(“ ་”)
  for str ∈ { ཀྱི, གྱི, གི, ཡི, དེ } **do**
    if tibetan_B.startswith(str) **then**
      tail ← tibetan_B[len(str) :]
    ▷ Replace the original function word with the new one, and add“ ་” again.
      ntb ← genitive_map(last_char())
      tokens[$i+1$] = ntb + tail + “ ་”
      break
    end if
  end for
end function
function Tibetan_sub(sent, old, new_word) tokens ← sent.split(“ ”)
  for $i \leftarrow 0$ **to** len(tokens) $- 1$ **do**
    if tokens[i] = old + “ ་” **then** tokens[i] ← new_word + “ ་”
      if $i + 1 <$ len(tokens) **then** func_word_adj(tokens, i)
      end if
      break
    end if
  end forreturn ¨. ¨join(tokens)
end function
text ← X
while true **do**
  Ti_sent ← T_RegexReplace(text, T)
  new_word ← R(Ti_sent) ▷ random replacement
  ▷ Reconnect the function words
  Tibetan_sen ← Tibetan_sub(Ti_sent, Ti_sent, new_word)
  $x_1 \leftarrow F$(Tibetan_sen)
  print(x_1)
end whilereturn x_1

Table 3 Tibetan Experimental Data

Data	Number of categories
TNCC	12
TU_SA	2
QHNUTSA	2
MZUCTSA	2

and text classification processing tasks. The TiBERT pre-trained model was proposed by the CMLI-NLP laboratory team of Minzu University of China to address the lack of native monolingual pre-trained models for Tibetan due to low resource availability [29]; the CINO pre-trained model was launched by the Harbin Institute of Technology-iFlytek Joint Laboratory (HFL) and is the first unified multilingual pre-trained model in China that systematically covers multiple ethnic languages [30], supporting 8 languages including Tibetan.

5.3 Evaluation indicators

In order to verify the effectiveness and applicability of the TWLASBMSL method in the morphological and syntactic linkage Tibetan word-level text attack scenario, this paper selects Accuracy Drop Value (ADV), Attack Success Rate (ASR), and Levenshtein Distance (LD) as core evaluation indicators in the research task of word-level adversarial sample generation. The Accuracy Degradation Value (ADV) represents the remainder of the total number of accurate original samples and the number of accurate models after word-level attacks. The specific calculation formula is as follows:

$$\mathrm{ADV} = \mathrm{Acc}_{\mathrm{original}} - \mathrm{Acc}_{\mathrm{adv}} \tag{1}$$

Where $\mathrm{Acc}_{\mathrm{original}}$ represents the accuracy of the original text, and $\mathrm{Acc}_{\mathrm{adv}}$ represents the model accuracy of the adversarial examples.

Attack Success Rate (ASR) is the quotient of the number of accurate samples after a word-level attack on the model, divided by the total number of attacked samples. The specific calculation formula is as follows:

$$\mathrm{ASR} = \frac{\mathrm{Acc}_{\mathrm{success}}}{\mathrm{Acc}_{\mathrm{total}}} \times 100\% \tag{2}$$

Here, $\mathrm{Acc}_{\mathrm{successl}}$ represents the total number of adversarial samples that were successfully attacked, and $\mathrm{Acc}_{\mathrm{total}}$ represents the total number of correct adversarial samples.

The Levenstein distance (LD) measures the difference between the original text and the adversarial example. The specific calculation formula is as follows:

$$\mathrm{LD} = \alpha \cdot \mathrm{Le}_{\max}(L_1, L_2) + T_p \tag{3}$$

In correcting the methods for adding core keywords and function words in Tibetan, the error detection rate, overall accuracy rate, and error correction success rate are used as the core evaluation indicators. The specific calculation formulas are as follows:

$$\mathrm{EDR} = \frac{N_{\mathrm{t_det}}}{N_{\mathrm{t_real_err}}} \tag{4}$$

Where $\mathrm{N}_{\mathrm{tdetl}}$ represents the number of Tibetan sentences with incorrect function words that the algorithm successfully identified, and $\mathrm{N}_{\mathrm{Ntrealerr}}$ represents the number of Tibetan sentences with actual errors when function words are added.

$$\mathrm{Acc} = \frac{N_{\mathrm{correct_after}}}{N_{\mathrm{total}}} \tag{5}$$

Where $\mathrm{N}_{\mathrm{correctafter}}$ represents the number of Tibetan sentences that are consistent with the standard sentence after the algorithm corrects the addition of Tibetan function words, and $\mathrm{N}_{\mathrm{total}}$ represents the total number of Tibetan sentences.

$$\mathrm{CSR} = \frac{N_{\mathrm{ti_fixed}}}{N_{\mathrm{ti_real_err}}} \tag{6}$$

Wherein, $\mathrm{N}_{\mathrm{tifixed}}$ represents the number of Tibetan sentences with erroneous function words that have been corrected to be accurate sentences, and $\mathrm{N}_{\mathrm{tirealerr}}$ represents the total number of Tibetan sentences with erroneous function words.

5.4 Experimental results

5.4.1 Word-level attack effect

To verify the feasibility of the word-level adversarial sample generation method proposed in this paper, this experiment uses a text adversarial method that combines morphology and syntax to conduct attack tests on the TiBERT and CINO-base-v2 and CINO-large pre-trained models on four test sets: TNCC, TU_SA, and QHNUTSA. The evaluation indicators of the experimental results are summarized in Tables 4 and 5 and Figure 2.

Table 4 Test results of the original Tibetan text

Model	data	Accuracy	Precision	Recall	F1
TiBERT	TNCC	66.92	64.16	62.51	62.77
	TU_SA	78.85	80.04	77.20	78.60
	QHNUTSA	72.50	74.67	68.10	71.23
	MZUCTSA	68.25	71.19	81.22	75.87
CINO-base-v2	TNCC	61.81	61.30	61.81	61.05
	TU_SA	67.39	72.54	69.17	74.82
	QHNUTSA	79.00	79.84	78.26	79.04
	MZUCTSA	77.42	76.01	74.85	75.31
CINO-large	TNCC	66.38	66.25	66.38	66.23
	TU_SA	76.53	75.13	77.60	74.35
	QHNUTSA	80.95	80.98	80.95	80.94
	MZUCTSA	80.14	79.02	79.73	79.31

Table 5 Test results after word-level adversarial attacks

Model	data	Accuracy	Precision	Recall	F1
TiBERT	TNCC	62.18	58.70	57.56	57.73
	TU_SA	76.75	78.58	74.10	76.28
	QHNUTSA	71.20	72.75	67.80	70.19
	MZUCTSA	68.93	71.16	83.15	76.69
CINO-base-v2	TNCC	58.58	58.28	58.58	57.76
	TU_SA	66.95	72.04	68.87	74.22
	QHNUTSA	76.60	77.93	75.00	76.44
	MZUCTSA	74.53	72.76	71.90	72.25
CINO-large	TNCC	64.12	64.47	64.12	64.16
	TU_SA	72.20	71.22	74.88	71.69
	QHNUTSA	79.60	79.63	79.60	79.59
	MZUCTSA	80.65	79.57	80.39	79.88

By comparing the experimental results of Accuracy, Precision, Recall, and F1 for the three pre-trained models TiBERT, CINO-base-v2, and CINO-large on the four text classification and sentiment analysis datasets TNCC, TU_SA, QHNUTSA, and MZUCTSA under original samples and attack samples in Tables 4 and 5 , it can be observed that the TWTADMBMI attack method leads to an overall decline in the classification performance of each model. The evaluation metrics of the three pre-trained models across different dataset combinations show varying degrees of decrease, and the resistance to attacks of different models on different Tibetan test datasets exhibits significant differences. Overall, the TWTADMBMI word-level adversarial attack results in significantly smaller reductions in metrics across all datasets, with the best robustness observed on the CINO-large model, while the overall attack resistance of the models is relatively weak.

Analysis of the above experimental results shows that all pre-trained target models, after being subjected to word-level adversarial attacks, experience a significant decline in recognition accuracy due to the attack method proposed in this paper. This indicates that the TWTADMBMI adversarial attack can effectively reduce model performance. Additionally, this attack causes the target models' ADV value to decrease by an average of 2.69%, with an average attack success rate (ASR) of 28.83%. The average edit distance (LD) between the original Tibetan text and the adversarial text is 6.80%. Although performance varies across different datasets, the attack method demonstrates outstanding overall performance in text sentiment analysis and classification tasks.

Figure 2 shows a relatively even distribution of ADV, ASR, and LD values. The attack not only generates stable and reliable adversarial examples on four different test datasets but also successfully interferes with the target model's normal predictions. Specifically, the ADV accuracy decrease reflects the following: On the TiBERT model, the TWTADMBMI attack method reduces the target model's ADV value by 1.30% to 4.74%, with the most significant effect on the TNCC test data, where the accuracy drops from 66.92% to 62.18%. On the CINO-base-v2 model, the ADV decrease ranges from 2.35% to 3.23%, slightly smaller than TiBERT, but the attack remains effective,

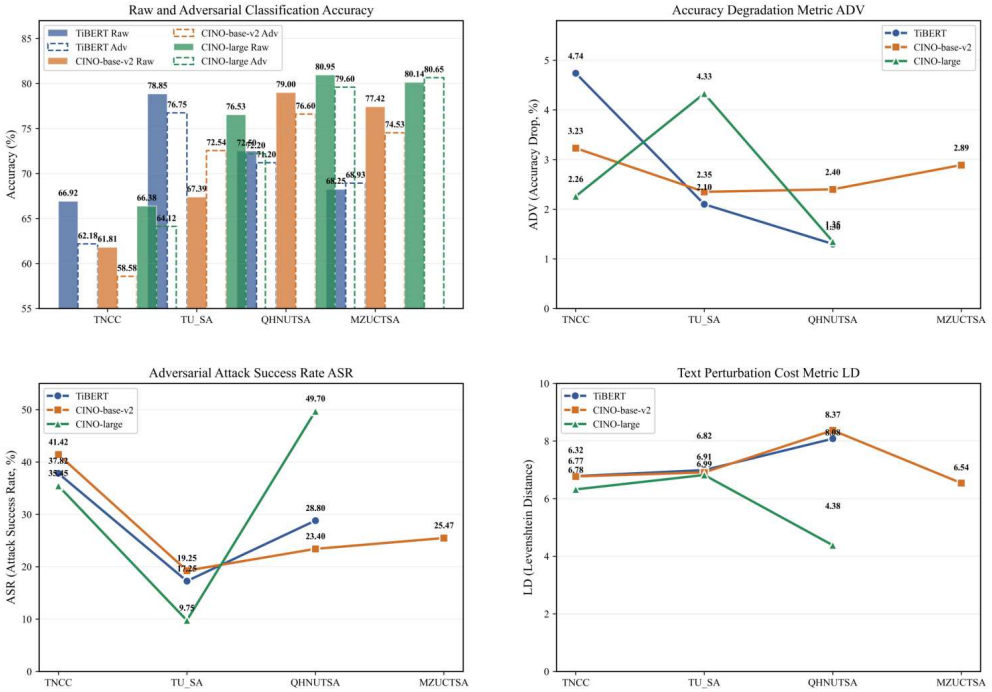


Fig. 2 Attack effects of word-level adversarial attacks on various datasets.
Note:Adversarial attacks against TiBERT and CINO-large models on the MZUCTSA dataset have not yielded effective results.

especially on the TNCC test dataset where the ADV value still decreases by 3.23%. On the CINO-large model, the ADV value decreases from 1.35% to 4.33%, showing a relatively stable attack effect.

The ASR attack success rates reflect that the CINO-large model achieves 49.70% on the QHNUTSA dataset, making it the best performing of the three models. Next is CINO-base-v2 with an ASR of 41.42% on the TNCC dataset, followed by TiBERT with 35.45% on the same dataset. Finally, neither TiBERT nor CINO-large achieved effective ASR on the MZUCTSA dataset. Overall, the TWTADMBMI attack method has the highest average ASR on the CINO-base-v2 model.

The Levinstein distance (LD) results indicate that the CINO-base-v2 model has the highest average LD value across various datasets, ranging from 6.54% to 8.37%, with a high value of 8.37% on the QHNUTSA dataset. TiBERT's average LD value falls between CINO-base-v2 and CINO-large, ranging from 6.78% to 8.08%. The CINO-large model has the lowest average LD value, ranging from 4.38% to 6.32%, indicating that the TWTADMBMI attack method achieves a high success rate with fewer word perturbations on the CINO-large target model, demonstrating optimal attack efficiency. The Levinstein distance similarity between the original Tibetan text

and the word-level adversarial examples is also closer. These LD results show that the TWTADMBMI method requires different word perturbation costs when attacking different target models, with the attack on CINO-base-v2 being the most difficult and the attack on CINO-large being the most efficient.

In summary, the proposed TWTADMBMI attack method demonstrates effective attack capabilities on four Tibetan test datasets and three pre-trained target models. It can precisely induce a reduction in the classification accuracy of the target model with small semantic perturbations, and the overall adversarial performance is reliable. Among them, the CINO-base-v2 model is the most sensitive to the TWTADMBMI attack method, exhibiting a high overall success rate and a more uniform distribution of generated perturbation amplitudes. CINO-large, on the other hand, shows the strongest robustness but is relatively more difficult to attack. Finally, experimental results fully validate the effectiveness and reliability of the TWTADMBMI method, a Tibetan word-level adversarial sample generation method based on morphosyntactic linkage, in adversarial attack tasks.

5.4.2 Analysis of the automatic adaptation and update of Tibetan function words

This paper leverages the advantages of the Tibetan Sememe Knowledge Base, Sememe Knowledge Base employing a random word perturbation method to replace the content of the original Tibetan text with words of higher semantic similarity through random matching and single-time replacement. Simultaneously, the replaced Tibetan function words are also replaced synchronously and consistently to ensure the generation effect and text quality of the Tibetan adversarial examples. On four test sets, the Tibetan function words following the replaced keywords were analyzed according to the rules for adding Tibetan function words. A judgment criterion was established for the automatic adaptation and updating of Tibetan function words after word perturbation replacement. Figure 3 below shows the experimental results of the dynamic adaptation and linkage changes when the replaced keywords are followed by additional words or used as the agent subject.

The data results in Figure 3 regarding the automatic adaptation and updating of Tibetan function words show that while random word replacement perturbation based on Tibetan Sememe Knowledge Base, through random matching of words with higher semantic similarity from the Tibetan Sememe Knowledge Base, and single word replacement, achieves the perturbation purpose, it also disrupts the grammatical logic rules of Tibetan text. Therefore, this paper selects the TNCC, TU_SA, QHNUTSA, and MZUCTSA Tibetan datasets and conducts experiments based on three Tibetan pre-trained models: TiBERT, CINO-base-v2, and CINO-large. For the original Tibetan text output by each model, based on the Tibetan Sememe Knowledge Base, the content in the original Tibetan text is randomly matched and replaced with words with higher semantic similarity in a single step. Finally, a Tibetan function word filling detection and correction algorithm is used to correct function words in the Tibetan text. The replaced Tibetan genitive, ergative, locative, and person noun suffix function words are also synchronously and consistently replaced, ensuring the generation effect and text quality of Tibetan adversarial examples.

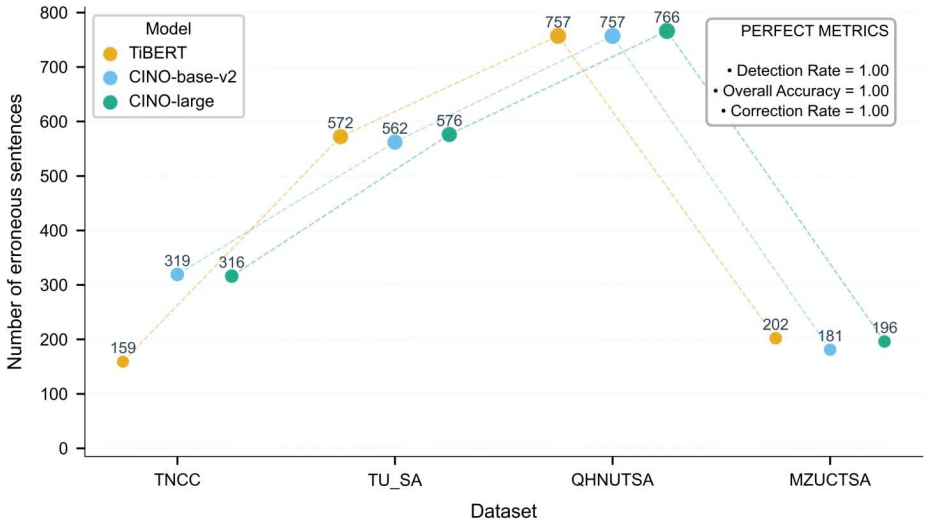


Fig. 3 Automatic adaptation and update/correction effect of Tibetan function words.

The sample size and the number of sentences with Tibetan function word errors vary across datasets. TiBERT uses 928 sample sentences from the TNCC Tibetan test data. CINO-base-v2 and CINO-large use 1854 and 1856 sample sentences respectively from the TNCC Tibetan test data. TU_SA and QHNUTSA both use 2000 Tibetan test sentences, while the MZUCTSA dataset uses a fixed 589. The detection and correction algorithm achieves 100% accuracy in all three metrics?error detection rate, overall accuracy, and complete error correction success rate?across all four datasets. It accurately locates function word errors in Tibetan test sentences, accurately identifies all test sample texts, and completes the correction of all erroneous sentences.

The above evaluation results fully demonstrate that the adversarial samples generated by the text adversarial attack strategy, which focuses on perturbing and replacing Tibetan character blocks, syllables, and words, contain a large number of errors in the attachment of Tibetan function words. This result clearly indicates that the replacement strategies for Tibetan character blocks, syllables, and words only achieve semantic perturbations at the lexical level, without truly constraining syntactic legitimacy. After minor formal modifications, the generated adversarial samples suffer from issues such as the failure of grammatical rules and broken sentence logic, leading to subpar quality.

In Tibetan, deep semantic relationships are realized through function words to organize sentences. To ensure the generation effectiveness and text quality of Tibetan adversarial examples, it is necessary to optimize the generation of word-level adversarial examples, then use a Tibetan function word addition detection and correction algorithm to complete the error correction of function words in Tibetan texts. Additionally, the replaced Tibetan genitive, ergative, locative, and personal noun suffix

function words are also consistently replaced synchronously, ensuring the generation effectiveness and text quality of Tibetan adversarial examples.

Table 6 provides several examples comparing the replacement of Tibetan word-level keywords with the addition of function words before and after the replacement.

Table 6 Examples of Tibetan word-level keyword replacement and function word re-addition

No.	Type	Tibetan Text	Category Shift
1	Original sample	མཚོ་སྨོན་པའི་ཅང་དུ་འབྱུག་གིས་ཆུ་འཁྱང་བའི་སྣང་ཚུལ་**རོ་མཚར་བ**་བྱུང་།	Positive→Negative
	Adversarial examples	མཚོ་སྨོན་པའི་ཅང་དུ་འབྱུག་གིས་ཆུ་འཁྱང་བའི་སྣང་ཚུལ་**ཡ་མཚན་ཞིག**་བྱུང་།	
	Translation	A peculiar hydrological phenomenon, resembling a waterspout, has been observed at Qinghai Lake.	
2	Original sample	ཁོ་ཚོས་**འཐབ་འཛིངས་**གི་བྱ་བར་གྱིས་བྱེད་བཞིན་ཡོད།	Politics→Economics
	Adversarial examples	ཁོ་ཚོས་**གཡུལ་འགྲེལ་**ཀྱི་བྱ་བར་གྱིས་བྱེད་བཞིན་ཡོད།	
	Translation	They are preparing for the battle.	
3	Original sample	དས་མོ་ལ་**དྲང་གཅང་**གི་དྲན་གདུང་དང་ལྡན་ཤེད་ཀྱི་འཆི་དུང་བཅིངས་ཡོད།	Literature→Politics
	Adversarial examples	དས་མོ་ལ་**ལྷག་བསམ་རོལ་ཤེད་**ཀྱི་དྲན་གདུང་དང་ལྡན་ཤེད་ཀྱི་འཆི་དུང་བཅིངས་ཡོད།	
	Translation	I have a sincere longing and a pure love for her.	
4	Original sample	སྐྱོ་སེམས་ཀྱི་རྟ་ཞོར་ཡིང་ཆེས་ཀྱི་གསོག་ཟུང་ཞིག་ཐོབ་ཅང་།	Literature→Environment
	Adversarial examples	དྲན་གདུང་གི་རྟ་ཞོར་ཡིང་ཆེས་ཀྱི་གསོག་ཟུང་ཞིག་ཐོབས་ངང་།	
	Translation	Longing is like a galloping horse; let it have wings so we can believe we will eventually meet again.	
5	Original sample	སྤྱན་སྔོན་ཀྱི་གནའ་རྟས་འདི་དག་ནི་རིམ་པ་དང་པོའི་རིག་རྟས་རེད།	Education→Economy
	Adversarial examples	མིག་མདུན་ཀྱི་གནའ་རྟས་འདི་དག་ནི་རིམ་པ་དང་པོའི་རིག་རྟས་རེད།	
	Translation	These antiques in front of us are first-class cultural relics.	
6	Original sample	གར་ཡོང་ཞི་ཤེས་པའི་དངོས་པོ་འདི་དག་གོང་རིམ་ནས་**གནང་བཤེས་**ཀྱི་རེ་བ་བཏོན་ཡོད།	Positive→Negative
	Adversarial examples	གར་ཡོང་ཞི་ཤེས་པའི་དངོས་པོ་འདི་དག་གོང་རིམ་ནས་**བསུ་ལེན་**ཀྱི་རེ་བ་བཏོན་ཡོད།	
	Translation	These items of dubious origin must be handed over by superiors.	

Note: Words marked in bold **orange** are randomly replaced based on the Tibetan semantic knowledge base; words marked in bold **green** are function words that need to be replaced; words marked in bold **red** are target words that have been replaced with the words that have the highest semantic similarity; and words marked in bold **blue** are function words that have been replaced with target keywords and then added back.

The test samples in the above table are generated based on the Tibetan semantic primitive knowledge base. By randomly matching and performing single-word replacements on the content from the original Tibetan text, they are substituted with words that have higher semantic similarity.and examples before and after random word replacement and function word reconnection in the generated adversarial samples, In adversarial example 1, " ངོམཚར་ " and " ཡ་མཚན་ " both mean strange things, but the target model changes from positive to negative when recognizing them. Only a small perturbation is needed to cause the model to misjudge. It is important to note that when the keyword " ངོམཚར་ " is replaced with the target keyword " ཡ་མཚན་ ", the following function word " བ " must also be changed to " ཞིག " to ensure the generation effect and text quality of the Tibetan adversarial example. Similarly, in adversarial sample 4, both " སྐྱོ་སེམས " and " དྭན་གངྡུང " mean sadness. When the target model identifies the text, it changes from literature to environment. The function word " ཀྱི " after the keyword also needs to be changed to " གི ". This can be considered a complete and high-quality Tibetan adversarial sample.

Overall, the keyword replacement, automatic adaptation of Tibetan function words, and update and concatenation rules proposed in this paper can better identify errors in function word collocation. All test datasets show a consistent trend of excellent performance in the experimental results, verifying the reliability of the method.

6 Conclusion

This paper, building upon existing research in Tibetan text adversarial attacks, addresses the issue of poor quality generated when using syllable- and word-level granular attack methods to produce adversarial samples. It proposes a novel morphosyntactic linkage adversarial sample generation method, TWTADMBMI, which conforms to Tibetan grammatical rules. This method leverages the advantages of the Tibetan Sememe Knowledge Base, using random word perturbation to randomly match and replace content in the original Tibetan text with words of higher semantic similarity. Simultaneously, it performs consistent replacement of the replaced Tibetan function words, ensuring both the quality and effectiveness of the generated adversarial samples. This avoids the problem of improper concatenation of Tibetan function words with core keywords. Experimental results demonstrate that the TWTADMBMI attack method exhibits superior performance in Tibetan sentiment analysis and text classification tasks, with better accuracy reduction, attack success rate, and Levenstein distance.

Due to limitations in the scope of the research and experimental conditions, this paper still has several areas that require improvement: Firstly, the Tibetan Sememe Knowledge Base constructed in this paper contains only 32,745 entries, with a limited corpus size and narrow semantic coverage, making it difficult to comprehensively collect vocabulary with similar semantic associations across multiple domains. Lastly, due to the scarcity and difficulty in using large models specifically designed for Tibetan, the target models employed in this paper are TiBERT, CINO-base-v2, and

CINO-large pre-trained models that support the Tibetan language. In the future, the TWTADMBMI attack method can be applied to Tibetan large models to effectively enhance their anti-interference and robustness. This issue holds significant potential for in-depth research.

Acknowledgements. This work is supported by the Science and Technology Major Project of the Science and Technology Program of the Xizang Autonomous Region (XZ202601ZD0010) and the Key Project of Xizang Natural Science Foundation (XZ202401ZR0040) .

Declarations. Conflict of interest The authors declare no competing interests.

References

[1] Zhao, W., Shun, C., Xiaoqiang, H.: Tibetan-chinese neural machine translation based on dynamic fusion of multilayer cino representations. Computer Engineering and Applications, 1–10 (2026). [J/OL]. Accessed 2026-04-27, https://link.cnki.net/urlid/11.2127.TP.20260302.0948.002

[2] Tashi, D., Chen, B., Sheng, T., Yu, Y., Wang, X., Zhang, J., Yeshi, L., Dongrub, R., Tsering, T., Tashi, N.: Tibetan data augmentation via gan-based handwritten text generation. CAAI Transactions on Intelligence Technology **11**(1), 55–65 (2026)

[3] Khysru, K., Tang, W., Wei, J., *et al.*: Research on speech synthesis technology based on tibetan rhythmic features. Expert Systems with Applications **277**, 127181 (2025)

[4] Qi, J., Wangdui, B., Jiang, J., Yang, J., Zhou, Y.: Edksanet: An efficient dual-kernel split attention neural network for the classification of tibetan medicinal materials. Electronics **12**(20), 4330 (2023)

[5] Miao, M.M., Kearns, M.: Hallucination, monofacts, and miscalibration: An empirical investigation. Proceedings of the National Academy of Sciences **123**(8), 2533582123 (2026)

[6] Kim, B.S., Kim, B., Jang, B.: Magnet: Counterfactual samples synthesizing for mitigating hallucination in large language models. Plos one **21**(2), 0340812 (2026)

[7] Yang, W., Yang, J., Guo, Y., Barthelemy, J.: Defensive dual masking for robust adversarial defense. Computational Linguistics, 1–38 (2025)

[8] Zhang, X., Hong, H., Hong, Y., Huang, P., Wang, B., Ba, Z., Ren, K.: Text-crs: A generalized certified robustness framework against textual adversarial attacks. In: 2024 IEEE Symposium on Security and Privacy (SP), pp. 2920–2938 (2024). IEEE

[9] El Lel, T., Ahsan, M., Latifi, M.: Lexicon-based random substitute and word-variant voting models for detecting textual adversarial attacks. Computers **14**(8), 315 (2025)

[10] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)

[11] Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2021–2031 (2017)

[12] Papernot, N., McDaniel, P., Swami, A., Harang, R.: Crafting adversarial input sequences for recurrent neural networks. In: MILCOM 2016-2016 IEEE Military Communications Conference, pp. 49–54 (2016). IEEE

[13] Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: Hotflip: White-box adversarial examples for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 31–36 (2018)

[14] Soffa, M.: Black-box generation of adversarial text sequences to evade deep learning classifiers. In: 2018 IEEE Security and Privacy Workshops (SPW) (2023)

[15] Jin, D., Jin, Z., Zhou, J.T., Szolovits, P.: Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8018–8025 (2020)

[16] Wang, T., Wang, X., Qin, Y., Packer, B., Li, K., Chen, J., Beutel, A., Chi, E.H.: Cat-gen: Improving robustness in nlp models via controlled adversarial text generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5141–5146 (2020)

[17] Zheng, X., Zeng, J., Zhou, Y., Hsieh, C.-J., Cheng, M., Huang, X.-J.: Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6600–6610 (2020)

[18] Zhang, S., Wu, H., Zhu, G., Xu, X., Su, M.: Character-level adversarial samples generation approach for chinese text classification. Journal of Electronics and Information Technology **45**(6), 2226–2235 (2023)

[19] Tong, X., Wang, L., Wang, R., *et al.*: Word-level adversarial sample generation method for chinese text classification. Information Network Security **20**(09), 12–16 (2020)

[20] Gong, Y., Zhang, X., Liu, Y., *et al.*: Adversarial sample generation method for

chinese text classification. Electronic Devices **46**(05), 1349–1356 (2023)

[21] Li, X., Luo, H., Sun, Y.: A chinese adversarial sample generation method based on chinese features. Journal of Software **34**(11), 5143–5161 (2023) https://doi.org/10.13328/j.cnki.jos.006744

[22] Wu, H., Li, X., Zhang, S., *et al.*: A method for generating chinese adversarial samples based on attack-guided diffusion. Journal of Tsinghua University (Natural Science Edition) **64**(12), 1997–2006 (2024) https://doi.org/10.16511/j.cnki.qhdxxb.2024.21.027

[23] Cao, X., Dawa, D., Qun, N., Nyima, T.: Pay attention to the robustness of chinese minority language models! syllable-level textual adversarial attack on tibetan script. In: Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), pp. 35–46 (2023)

[24] Cao, X., Qun, N., Gesang, Q., Zhu, Y., Nyima, T.: Multi-granularity tibetan textual adversarial attack method based on masked language model. In: Companion Proceedings of the ACM Web Conference 2024, pp. 1672–1680 (2024)

[25] Cao, X., Gesang, Q., Sun, Y., Qun, N., Nyima, T.: Tscheater: Generating high-quality tibetan adversarial texts via visual similarity. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2025). IEEE

[26] Li, J., Cao, X., Qun, N., *et al.*: Tibetan adversarial sample generation for classification tasks. Plateau Scientific Research **9**(04), 118–128 (2025) https://doi.org/10.16249/j.cnki.2096-4617.2025.04.012

[27] Du, X., Wu, H., Yi, Z., *et al.*: A review of text adversarial sample attack and defense techniques. Journal of Chinese Information Processing **35**(08), 1–15 (2021)

[28] Caihua: A study on the syntactic and semantic integration of tibetan syntax based on case grammar. PhD thesis, Tibet University (2018)

[29] Liu, S., Deng, J., Sun, Y., Zhao, X.: Tibert: Tibetan pre-trained language model. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2956–2961 (2022). IEEE

[30] Yang, Z., Xu, Z., Cui, Y., Wang, B., Lin, M., Wu, D., Chen, Z.: Cino: A chinese minority pre-trained language model. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3937–3949 (2022)