Environment and Social Psychology (2025) Volume 10 Issue 7 ISSN: 2424-8975 (O)
doi: 10.59429/esp.v10i7.3834 2424-7979 (P)

RESEARCH ARTICLE

Generating culturally-contextual Chinese cyberbullying datasets: A GAN approach for social psychology research

Xiangqin Dai[1,2], Mohd Najwadi Yusoff[1], Bingli Zhu[2], Xiao Zhang[1], Wujin Jiang[2], Lei Wang[1]

[1] *School of Computer Sciences, Universiti Sains Malaysia, Pulau, Pinang, 11800, Malaysia*

[2] *Key Laboratory of Intelligent Information Processing and Control, Chongqing Three Gorges University, Chongqing, 404000, China*

*** Corresponding author:** Mohd Najwadi Yusoff, najwadi@usm.my

ABSTRACT

Cyberbullying has become a growing concern with serious psychological and social consequences, including anxiety, depression, and disrupted online communities. Grounded in social psychology theories such as social learning and online disinhibition, cyberbullying is shaped by factors like anonymity and peer influence. However, the lack of Chinese-language cyberbullying datasets limits research and intervention efforts. To address this, we used four GAN models SeqGAN, RankGAN, MaliGAN, and LeakGAN to generate realistic Chinese cyberbullying text. LeakGAN outperformed the others, achieving a BLEU2 score of 0.948, self-BLEU2 of 0.963, NLL of 0.48, and the highest EmbSim values. Beyond technical performance, we emphasized psychological validity, cultural relevance, and ethical considerations in the data generation process. The findings have important implications for automated detection, intervention design, and social psychology research. Framed within ecological systems theory, this work also considers how online environments shape behavior. The synthetic dataset supports applications in schools, workplaces, and cross-cultural studies, though limitations remain in capturing the full complexity of real human behavior. Overall, LeakGAN's success offers a strong foundation for future research on cyberbullying in digital contexts.

Keywords: generative adversarial networks; Chinese cyberbullying dataset; LeakGAN; cyberbullying

1. Introduction

Cyberbullying is defined as "aggressive and deliberate behavior, repeated over time, using electronic communication means by a group or individual against victims unable to defend themselves." This highlights the persistent harm caused through online harassment, defamation, and threats. Research shows that cyberbullying has become a significant global concern, particularly among adolescents. With the rapid rise of digital technologies, youth who are still undergoing psychological development have become especially vulnerable. Studies have shown that cyberbullying can severely impact minors ' mental and physical health, leading to academic difficulties, social withdrawal, anxiety, depression, and even post-traumatic stress symptoms. A recent survey by China Youth Daily indicated that over 65% of respondents had experienced or witnessed online violence, and nearly 72% observed an upward trend in such

ARTICLE INFO
Received: 23 June 2025| Accepted: 26 June 2025 | Available online: 10 July 2025

CITATION
Dai XQ, Yusoff MN, Wang L ,et.al. Generating culturally-contextual chinese cyberbullying datasets: a gan approach for social psychology research. *Environment and Social Psychology* 2025; 10(7): 3834 doi:10.59429/esp.v10i7.3834

incidents.

From a social psychological perspective, theories such as social learning, online disinhibition, and deindividuation explain how anonymity and reduced accountability foster harmful behavior in digital spaces. Cyberbullying disrupts both individual well-being and the broader social environment, undermining trust, social cohesion, and psychological safety in online communities. Moreover, the ecological systems theory offers a useful lens to examine how online platforms, peer groups, families, and societal norms interact to either enable or suppress cyberbullying.

Although institutions ranging from governments to NGOs have implemented strategies to combat online abuse, effective intervention and prevention rely on the availability of high-quality, annotated datasets. However, current Chinese cyberbullying datasets suffer from four major limitations: (1) Insufficient Quantity existing datasets are too small to effectively train deep learning models; (2) Inaccurate Annotation subjectivity and inconsistency reduce training effectiveness; (3) Lack of Diversity data is often platform-specific and fails to represent varied cyberbullying behaviors; and (4) Outdated Content evolving digital language and trends are not reflected in older datasets, weakening model relevance.

To overcome these limitations, this study proposes the use of Generative Adversarial Networks (GANs) specifically, models like SeqGAN, RankGAN, MaliGAN, and LeakGAN to synthetically generate high-quality Chinese cyberbullying datasets. These models are capable of producing large-scale, linguistically diverse, and psychologically relevant data. Special emphasis is placed on ensuring the psychological validity of the generated content such as realistic emotional tone, contextual fit, and aggression patterns which is critical for use in behavioral and clinical studies. The data generation process also prioritizes cultural and linguistic alignment, accounting for regional idioms, online slang, and communication norms specific to Chinese digital platforms.

Ethical considerations are central to the methodology. The generation and use of potentially harmful content are carefully reviewed through the lens of psychological ethics, including risk of desensitization, potential misuse, and psychological harm to researchers or annotators. The project adopts safeguards such as content filters, ethical review processes, and expert validation to uphold research integrity.

The practical implications of this research are significant. The enriched datasets can enhance the accuracy of AI-based detection systems, support early warning interventions, and inform social media moderation policies. In social psychology research, the datasets allow for simulations and studies of group dynamics, aggression escalation, and emotional contagion in digital spaces. Through the lens of environmental psychology, the work sheds light on how digital design, interface structure, and perceived safety influence user behavior and mental health.

Beyond academic contributions, the datasets have strong potential for real-world applications. In educational settings, they can be used to train automated monitoring tools or guide teacher interventions. In workplaces, they can inform HR policies against online harassment. Additionally, this approach allows for cross-cultural comparison of cyberbullying patterns, promoting international collaboration on digital safety. Still, caution is warranted: synthetic datasets, while useful, may not fully capture the nuanced emotional experience of real-life victims or aggressors. Future research should incorporate clinical validation and user feedback to bridge this gap.

In conclusion, this study offers a multi-disciplinary contribution by combining advanced machine learning with psychological theory and ethical research practice. By leveraging GANs to generate

culturally and psychologically grounded datasets, it lays a strong foundation for more effective detection, intervention, and understanding of cyberbullying in Chinese digital environments.

2. Related work

Generative Adversarial Networks (GANs) have emerged as a powerful tool across domains such as image enhancement, medical imaging, style transfer, and increasingly, natural language generation, where they enable the creation of high-quality, diverse, and contextually appropriate text. In the realm of text generation, GAN-based models such as SeqGAN[23], LeakGAN[24], and CCGAN[25] have made significant progress.

SeqGAN introduced reinforcement learning to treat text generation as a sequential decision-making process, effectively addressing the challenge of non-differentiability in discrete text data. LeakGAN expanded upon this by incorporating a hierarchical generator and supervisory leakage from the discriminator to improve syntactic structure and content relevance. CCGAN leveraged specific features of the Chinese language such as tone sensitivity and character-level context to enhance linguistic fluency and semantic consistency. More recently, TILGAN[26] and FGGAN[27] employed transformer architectures and feature-guided generation modules, respectively, to improve semantic control and syntactic coherence, while SAL-GAN[35] addressed sparse reward and mode collapse problems with self-adversarial learning.

Despite these advancements, GAN applications to cyberbullying datasets remain underdeveloped, especially for Chinese-language content. This study breaks new ground by being the first to apply GAN to generate Chinese cyberbullying text, with three primary objectives: (1) to expand the scale of available data to support deep learning-based cyberbullying detection; (2) to improve dataset diversity across platforms and bullying types; and (3) to retain cultural and emotional realism consistent with actual online interactions in Chinese digital environments.

From a social psychology perspective, the generated content is designed to reflect theoretical mechanisms underpinning cyberbullying behavior. For example, Bandura's Social Learning Theory explains how users imitate harmful behavior observed in online spaces, particularly when such behavior is rewarded such as with likes or attention. The Online Disinhibition Effect (Suler, 2004) helps explain why individuals behave more aggressively online due to perceived anonymity, reduced empathy, and asynchronous feedback. Furthermore, deindividuation theory and social identity theory explain how group dynamics on platforms like QQ groups, WeChat communities, or Douyin comment sections amplify aggression through depersonalization and in-group/out-group bias. GAN-generated datasets that simulate these dynamics can be used to explore how various social cues emoji use, slang, collective commenting contribute to hostile discourse.

In terms of environmental context, this study considers platform-specific characteristics that influence the form and frequency of cyberbullying. For instance, on Weibo, public visibility and algorithmic amplification can escalate harassment campaigns such as mass retweeting abusive content. On WeChat, bullying is more private and relational, involving exclusion from groups or spreading false rumors. Douyin (TikTok China) fosters harassment through video comment chains and duets that ridicule others. Using Bronfenbrenner's Ecological Systems Theory, cyberbullying is modeled as a multi-layered phenomenon:

Microsystem: peer interactions via chat apps or comment sections;

Mesosystem: the interplay between online and offline environments (e.g. school bullying extending into digital spaces);

Exosystem: platform algorithms, content moderation policies;

Macrosystem: societal norms, media portrayals of aggression, and cultural acceptance of "cyber shaming."

Methodologically, the GAN-generated text is not only evaluated for linguistic coherence but also psychological authenticity. This includes affective analysis to verify emotional tone such as hostility, sarcasm, pragmatic labeling of social function such as exclusion, humiliation, and context validation by native speakers familiar with Chinese digital discourse. Culturally specific phrases like "键盘侠"(keyboard warrior) or "人肉搜索" (doxxing) are incorporated to ensure contextual relevance. Moreover, manual annotations and expert psychological review are employed to validate that generated content matches the emotional and behavioral profiles seen in real cyberbullying incidents.

Ethical considerations are addressed throughout. To minimize harm, synthetic text is filtered for extreme content and presented in anonymized, non-targeted form. All human annotators undergo psychological briefing and have the option to opt out of emotionally disturbing content. The research complies with APA ethical guidelines and IRB standards regarding exposure to emotionally sensitive material.

In terms of discussion and application, the study offers robust implications for both psychological research and real-world interventions:

In school environments, GAN-generated scenarios can be used to train AI-driven monitoring tools or inform educator-led intervention simulations;

In workplaces, the data helps develop HR screening systems for digital harassment and psychological safety audits;

In therapeutic settings, psychologists can use synthetic text as case material to help clients understand, role-play, and navigate cyberbullying episodes.

From a research perspective, the dataset supports computational social psychology experiments on digital aggression, environmental psychology studies on how platform design affects user behavior, and cross-cultural psychology by comparing aggression patterns in Chinese versus Western digital discourse. Furthermore, it enables comparative research on gendered cyberbullying language, bystander intervention modeling, and linguistic toxicity escalation.

However, the study acknowledges key limitations. While the generated content mimics surface-level language patterns, it may not capture the full emotional nuance, intent, or psychological complexity of real user interactions. GANs cannot currently replicate lived experiences, trauma narratives, or contextual subtext with full accuracy. Therefore, synthetic data should be seen as a complement to, not a replacement for, real-world data, and should be validated with human-in-the-loop protocols in future studies.

In summary, this study advances both technological and psychological frontiers by using GANs to generate culturally grounded, emotionally authentic, and socially relevant Chinese cyberbullying datasets. It supports practical interventions, enriches social psychological understanding of online aggression, and provides scalable, ethical tools for combating cyberbullying in today's digital ecosystems.

3. Methods

3.1. Research frame

This study addresses the scarcity of Chinese cyberbullying datasets, which has significantly constrained the advancement of Chinese cyberbullying detection technologies. By leveraging Generative Adversarial Network (GAN) models to generate high-quality Chinese cyberbullying datasets, the study aims to provide more diverse and realistic training data for cyberbullying detection algorithms, thereby enhancing the performance of detection models.

The primary objective of this research is to explore and implement the generation of Chinese cyberbullying datasets based on GAN models. The goal is to produce high-quality and diverse datasets that meet the demands of research and practical applications.

In this study, the process is divided into four key steps: data collection, data processing, GAN model training, and data evaluation, as illustrated in **Figure 1**.

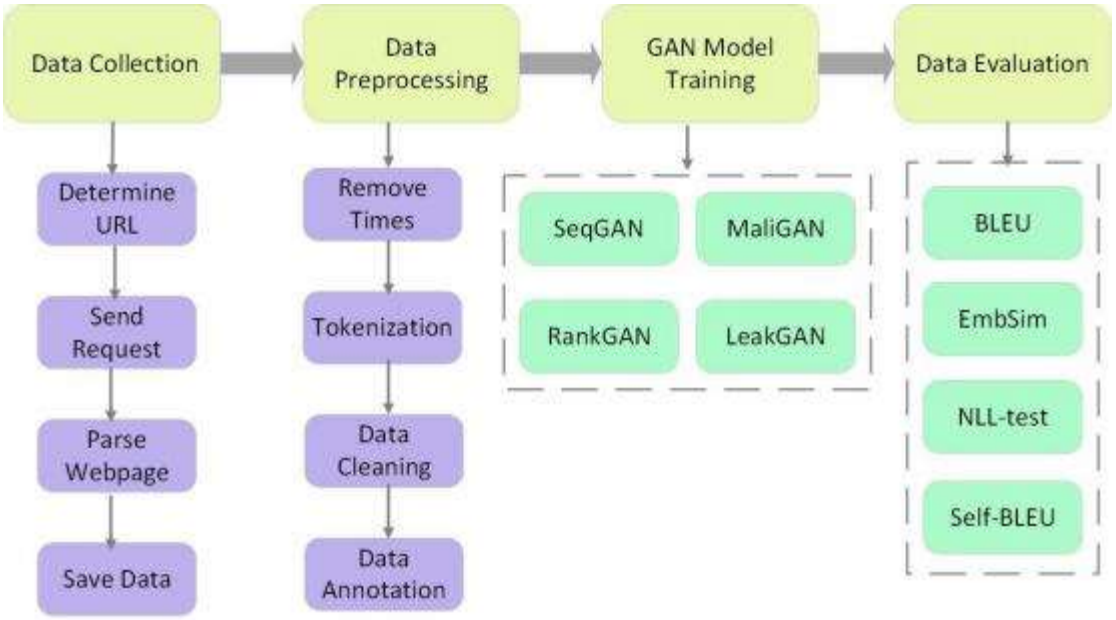

Figure 1. Research methodology.

Step 1: Data Collection

The first step involves determining the target URLs of Weibo posts, sending requests, and receiving responses. Subsequently, the page content is parsed and the data is saved. The collected data includes comment timestamps and comment content.

Step 2: Data Preprocessing

In this step, the comment timestamps are removed, and tokenization is performed using Jieba. Stop words are eliminated, followed by data cleaning to remove noise and filtering out low-frequency words. Word vectors are then constructed using Word2Vec. Finally, the dataset is labeled. Each comment in the dataset is independently evaluated by three reviewers to determine the presence of cyberbullying. Text content is first filtered, and comments deemed to contain cyberbullying by at least two reviewers are selected. A total of 4,287 comments were collected, as shown in Figure 2, to serve as the training data for the model.

Step 3: GAN Model Training

Four different GAN models were utilized for training. Detailed descriptions of these models are provided in Section 3.2.

Step 4: Data Evaluation

The evaluation process employs four evaluation metrics. Detailed explanations are presented in Section 3.3.

June 10th 23:43 Strongly request that three identical heroes do not need to be automatically synthesized! I want to put a row of Yasuo! I kowtow ! ! ! ! 0 0 0
June 10th 23:44 Hero battle flag mode??! I'm already looking forward to it 0 0 0
June 10th 23:44 Do it! 0 0 0
June 10th 23:49 Old players like me are blessed, playing chess and slacking off at work every day 0 0 0
June 10th 23:44 The sooner it goes online, the better, please speed up. 0 0 0
June 10th 23:48 Looks very interesting 0 0 0
June 10th 23:54 Start looking forward to it [applause] 0 0 0
Today 00:04 I took a look and it's not a mode I can understand 0 0 0
June 10th 23:51 The official blog posts messages slower than I do 0 0 0
Today 00:14 Go go go, looking forward to it 0 0 0
June 10 23:56 LOL Auto Chess 0 0 0
Today 13:56 Really low 1 1 1
Today 00.04 Auto Chess also started copying? Can you, a bankrupt client, figure it out? 1 1 1
Today 00:01 Copying started? 1 1 1
June 10 23:54 Start looking forward to it 0 0 0
June 10 23:43 LOL Auto Chess 0 0 0
June 10 23:43 When everyone thought that the last battle must be a bayonet fight between Giant Bird Duoduo and Valve, Tencent stood up with two rocket launchers.··· (hot review stolen from the next door) 0 0 0
June 10 23:46 Those who say it's stolen, I suggest you sue Riot [Quan] in court. China United Network sued, I support you. If I didn't sue you, why are you saying you're a jerk? Ning said, is that right? 1 1 1
June 10 23:45 Dota2 Auto Chess doesn't understand heroes, but League of Legends will understand, hahaha, go ! ! ! 0 0 0
June 10 23:44 If you open your own map, I can still look up to you. Plagiarism, ugh. Tencent is really good. 1 1 1

Figure 2. Weibo dataset.

3.2. Model introduction

Four text generation models based on GAN are employed: SeqGAN[37], MaliGAN[38], RankGAN[39], and LeakGAN[40]. These models encompass supervised and unsupervised methods, adversarial techniques, and hierarchical approaches.

SeqGAN as shown in **Figure 3**,utilizes a discriminator model trained to minimize the binary classification loss between real and generated texts. Additionally, the generator, following a pre-training process based on Maximum Likelihood Estimation (MLE), employs the REINFORCE algorithm to optimize the GAN objective. The formula is as follows:

$$\min_{\varphi} - \mathbb{E}_{Y \sim p_{\text{data}}} \left[\log D_{\varphi}(Y) \right] - \mathbb{E}_{Y \sim G_{\theta}} \left[\log(1 - D_{\varphi}(Y)) \right]. \tag{1}$$

φ is a parameter of discriminator D_{φ}, $\mathbb{E}_{Y \sim p_{\text{data}}}$ denotes the expected value calculated after sampling data Y from the real data distribution pd_{ata}, $\mathbb{E}_{Y \sim G_{\theta}}$ represents the expected value obtained after sampling data Y from the generator G_{θ}.

To reduce variance, SeqGAN utilizes a Monte Carlo search to calculate the Q-values for each generated token. MaliGAN, adhering to the fundamental framework of SeqGAN, readjusts rewards within a batch size of mmm to stabilize training and alleviate gradient saturation issues.

$$r_D(x_i) = \frac{r_D(x_i)}{\sum_{j=1}^{m} r_D(x_j)} - b \ , \tag{2}$$

The reward function $r_D(.)$ in MaliGAN is derived from the discriminator, representing the moving average as a baseline. RankGAN[39] replaces the discriminator in SeqGAN with a ranker, optimizing the ranking loss.

$$L_{\varphi} = \mathbb{E}_{s \sim p_{\text{data}}} \left[\log R_{\varphi}(s \mid U, C^{-}) \right] - \mathbb{E}_{s \sim G_{\theta}} \left[\log R_{\varphi}(s \mid U, C^{+}) \right], \tag{3}$$

Where

$$R_{\phi}(s \mid U, C) = \log \left(\frac{\exp(\gamma \alpha(s \mid u))}{\sum_{s' \in C} \exp(\gamma \alpha(s' \mid u))} \right)$$
$$\alpha(s \mid u) = \cos(y_s, y_u) \ , \tag{4}$$

LeakGAN is a hierarchical reinforcement learning framework consisting of two modules, termed the "Manager" and the "Worker." The Manager typically learns to set a series of sub-goals for sequence generation, while the Worker learns to achieve these sub-goals.

The manager captures global features by learning the long-range dependencies within the entire text sequence, generating vectors to guide the generation process. The formulas for feature extraction and target vector generation are as follows:

Extract features f_t from the partially generated sequence $S_{1:t}$

$$f_t = CNN(S_{1:t}) \ , \tag{5}$$

The manager generates the target vector g_t based on f_t

$$g_t = W_g f_t + b_g \ , \tag{6}$$

$W_g f_t$ is an linear transformation after feature extraction. b_g is an flexibility to the linear transformation.

LeakGAN's generator uses a Long Short-Term Memory Network (LSTM), and the discriminator usually also employs a Convolutional Neural Network (CNN), which generates a word at each step and adapts the generation strategy by combining the leaked feature information.

3.3. Evaluation metrics

This research employs four evaluation metrics: BLEU[42], EmbSim[41], NLL-test[41], and Self-BLEU[41]. BLEU and EmbSim are document similarity-based metrics, NLL-test is a likelihood-based metric, and Self-BLEU is a diversity-based metric.

3.3.1. Document similarity-based metrics

The most direct way to assess the quality of generated documents is by measuring their similarity to the natural language or the training dataset. BLEU is a widely used metric for evaluating the word-level similarity between sentences or documents. EmbSim, inspired by BLEU, was proposed to evaluate the similarity between two documents. EmbSim stands for "embedding similarity" and instead of comparing sentences word by word, it compares the embedding representations of words.

First, the skip-gram model is used to evaluate word embeddings on real data. For each word embedding, the cosine distance to other words is \ calculated and then structured into a matrix W , Here, $Wi,j = \cos(ei, ej)$, ei and ej are the word embeddings of words i and j from the real data.

W represents the similarity matrix derived from the real data.

Similarly, the similarity matrix W' for the generated data is obtained.

Here, $W'_{i,j} = \cos(e'_i, e'_j)$ and e'_i , e'_j are the word embeddings of words derived from the generated data using the same skip-gram model.EmbSim is defined as:

$$\text{EmbSim} = \log\left(\sum_{i=1}^{N} \cos(W_i', W_i) / N\right) \tag{7}$$

Where N is the total number of words, W_i and represents the i -column of the similarity matrices W and W' , respectively.

The goal of MLE is to minimize the cross-entropy between the real data distribution p and the model-generated data distribution q. Metrics can be designed to assess how well the data fits the model by measuring the likelihood. These metrics require detailed information not only about the data but also about the model itself.

Negative log likelihood (NLL) was initially introduced in SeqGAN[37], specifically for synthetic data experiments, to measure how well the generated data fits a reference language model. In the NLL-oracle approach, a randomly initialized Long Short-Term Memory (LSTM) network is considered as the reference model, or "oracle." The text generation model aims to minimize the average negative log-likelihood of the generated data concerning the oracle LSTM. that is,

$i_{x \sim q}\left[\log p(x)\right]$ where x denotes the generated data

Since the LSTM is treated as the reference model, this metric allows for the computation of the average loss on a per-sentence and per-word basis.

$$NLL_{oracle} = -i_{Y1:T \sim G}\left[\sum_{t=1}^{T} \log G_{oracle}\left(Y_t \mid Y_{1:t-1}\right)\right], \tag{8}$$

where Goracle denotes the oracle LSTM, and Gθ denotes the generative model.

NLL-test is a straightforward metric used to evaluate the model's ability to adapt to real test data, and it corresponds to NLL_{test}

$$NLL_{test} = -i_{y \sim G_{real}}\left[\sum_{t=1}^{T} \log G_{\theta}\left(y_t \mid Y_{1:t-1}\right)\right], \tag{9}$$

where G_{real} represents the distribution of the real data.

NLL-test applies only to autoregressive generators such as RNNs, $G\theta\ (yt \mid Y1:t-1)$ given a generator, it requires the likelihood of a specific word to be computed based on the preceding words.

3.3.3. Diversity metrics

GAN is often plagued by the mode collapse issue, where the generator produces a limited number of samples or only a single type of hig similar samples. Consequently, in open-domain text generation tasks, metrics that encourage the generation of more diverse patterns are included.

Self-BLEU[41] is a metric used to assess the diversity of generated data. While BLEU is typically employed to evaluate the similarity between two sentences, it can also be applied to measure the similarity of a sentence within a set of generated sentences to the other sentences in that set. By treating one sentence as the hypothesis and the remaining sentences as references, BLEU scores can be calculated for each generated sentence, and the average BLEU score is defined as the Self-BLEU score for the document. A higher Self-BLEU score indicates lower diversity within the document and signifies more severe mode collapse in the GAN model.

4. Experiments and results

4.1. Generative adversarial network (GAN) training strategy

The training strategy consists of three steps, as detailed in Flowchart 3 and the algorithm pseudocode shown in **Table 1**

1. Initialization Phase

All generators are initialized with parameters sampled from a Gaussian distribution N (0, 1), it represents a normal distribution with a mean of 0 and a standard deviation of 1, used to randomly initialize the generator's weights.

2. Maximum Likelihood Estimation (MLE) Pretraining Phase

Generator Training: The generator is pretrained for 80 epochs using the maximum likelihood estimation (MLE) method. During this phase, the generator's parameters are adjusted to ensure the sequences it generates closely approximate the real data distribution.

Discriminator Training: The discriminator is also pretrained for 80 epochs. It learns to differentiate between real data and the generated data, improving its ability to identify generated samples.

3. Adversarial Training Phase

Generator Update: During each adversarial training cycle, the generator is updated once based on feedback from the discriminator.

Discriminator Update: Following each generator update, the discriminator undergoes 15 mini-batch gradient updates to strengthen its ability to discriminate between real and generated samples effectively.

Periodic MLE Training: After every 10 adversarial training cycles, both the generator and discriminator undergo 5 epochs of additional MLE training. This step helps maintain model stability, preventing issues like overfitting or mode collapse commonly associated with adversarial training.

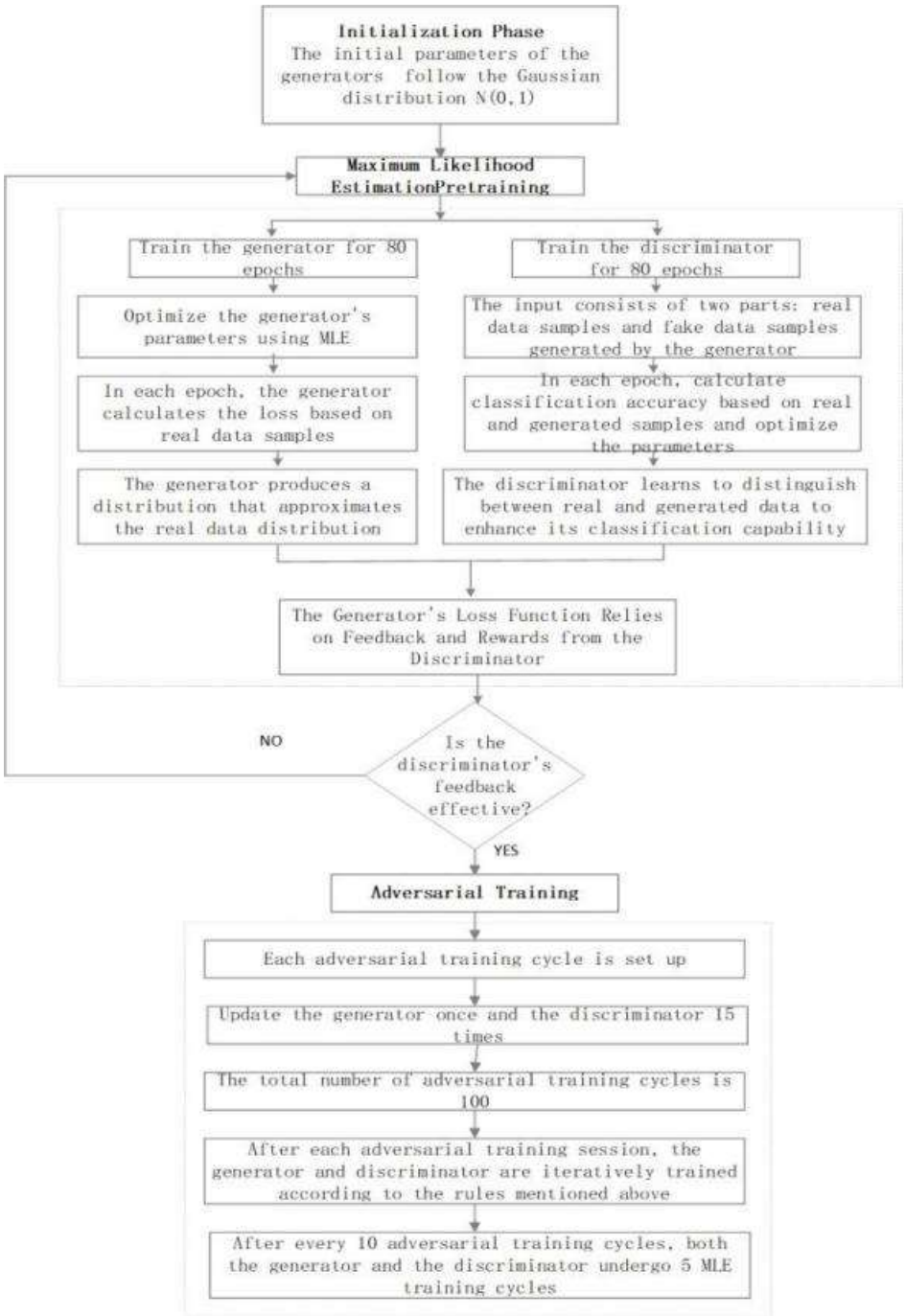

Figure 3. Flowchart oftraining strategy.

Table 1. Algorithm pseudocode.

```
Algorithm: Adversarial Training with MLE Pre-training

Input: Generator G, Discriminator D, Total adversarial cycles T = 100,
MLE pre-training epochs E_mle = 80,
        Adversarial discriminator updates D_updates = 15, MLE refresh
cycles R_cycle = 10,
        MLE refresh epochs R_epochs = 5

Output: Trained generator G and discriminator D

Initialize parameters of G and D with Gaussian distribution N(0, 1)
# Pre-training with Maximum Likelihood Estimation (MLE)

for epoch in range(E_mle):

    Train G using MLE

for epoch in range(E_mle):
    Train D using MLE

# Adversarial training
for cycle in range(T):

    # Update G once

    Train G adversarially using D


# Perform D_updates mini-batch gradient updates for D
for _ in range(D_updates):
                Train D using G

# Every R_cycle adversarial cycle, perform additional MLE training
            if (cycle + 1) % R_cycle == 0:

                for epoch in range(R_epochs):

                    Train G using MLE

                for epoch in range(R_epochs):

                    Train D using MLE
    return Trained generator G and discriminator D
```

4.2. Experimental results

Generated cyberbullying dataset and example of generating samples from a real dataset as shown in **Figure 4** and **Figure 5**

Remember this head stinky eye can not be used as
No shame
You are really worse than the scalpers who are shameless. They go to such people and don't even apologize when they come out!!!!!
Where should you use it? Shameless, shameless, you even
People like you are a bit crazy!!!!!
Are you a bitch!!!!!
You are such a bitch, you are a woman, there is no problem with so many people, you woman, we are you, you are worthy???????????????
Follow your horse talk
You and your father, shameless
Is there anyone who doesn't know why you don't die?
You really can't forgive!
The tail of the dog is desperate?
Shameless, still want to die
You are just a coward?
Shameless?
Fuck you Yuntian, this person is really to username Your mother
Excuse me, are you Damai Netcom!!!!!
You rely on your own country's violent express delivery, do you really think you are something, do we mean it alone? Are you alone?
Is it also true that you buy scalpers online? A person, shameless, has been talking to you me, you are shameless, can
You are not worthy of being a human being, garbage
It's so ugly
Your airport speech brain is simply shameless, you deserve to be sent
Am I a mad dog??????????????????????????????????
Did the blue and red words take turns?

Figure 4. Generated cyberbullying dataset.

DATASETS	Cyberbullying
Test Set	Heaven cursed the black dog
	Take it to court, China Securities Regulatory Commission, I support you. If you don't sue me, why are you saying you are a bitch?
SeqGAN	God cursed the black dog for not being worthy to be here
	Take it to court, China Securities Regulatory Commission sued it. I admit that you didn't sue me. Why are you saying that you are a bitch?
SeqGAN	God cursed the black dog for being sick
	Take it to court. China United Network sued me. I support you. If you didn't sue me, you're here saying you're a bitch.
RankGAN	God damn the yellow dog is so shameless
	Take it to court. China Securities Regulatory Commission sued me. I accept that you didn't sue me. What the hell are you talking about here?
RankGAN	God curses the yellow dog, get out
	Take it to court. China Securities Regulatory Commission sued me. I support you. If you don't sue me, why are you here talking about you?

Figure 5. Example of generating samples from a real dataset.

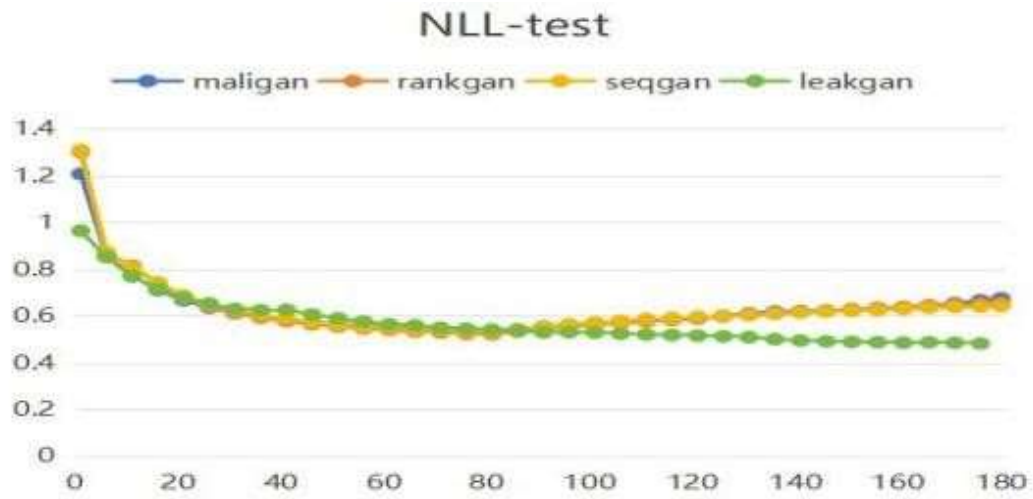

Figure 6. NLL-test.

From **Figure 6**, it can be observed that the NLL values for all models gradually decrease as the adversarial training epochs increase. This indicates that the performance of all models in generating cyberbullying text improves over time, with the generated text increasingly resembling the distribution of real data. LeakGAN consistently exhibits the lowest NLL values at all time points, indicating that it achieves the highest likelihood of generating cyberbullying text, meaning it produces text that is closer to the real data. RankGAN and MaliGAN also show a downward trend in NLL values during training, suggesting that they may have certain biases or inadequacies in text generation. SeqGAN starts with relatively high NLL values, but these values gradually decrease as training progresses. However, throughout the training process, SeqGAN's NLL values remain higher than those of the other three models, indicating that SeqGAN's performance in generating cyberbullying text is relatively weaker.

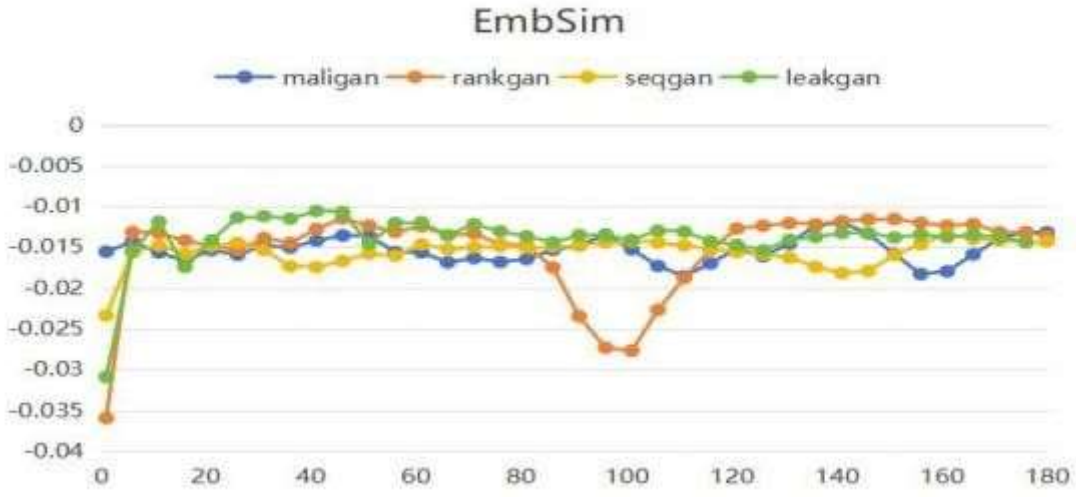

Figure 7. EmbSim.

From **Figure 7**, The effectiveness of adversarial training for the four models suggests that MLE training cycles after every 10 adversarial epochs may help stabilize model performance and improve the quality of generated text. This is evident from the overall upward trend in EmbSim values, which, despite some fluctuations, generally move towards higher similarity.

In the task of generating cyberbullying text, LeakGAN consistently exhibits the highest EmbSim values at most time points, indicating its optimal performance in generating cyberbullying text that closely aligns with the semantic characteristics of the target text. In contrast, SeqGAN shows more stable performance but has relatively lower EmbSim values, suggesting limitations in generating cyberbullying text that is highly similar to the target text. Additionally, the performance of RankGAN and MaliGAN on the EmbSim metric falls between SeqGAN and LeakGAN, demonstrating distinct characteristics in their text generation capabilities.

Table 2. BLEU.

	SeqGAN	**MaliGAN**	**RankGAN**	**LeakGAN**
BLEU-2	0.774	0.779	0.804	0.948
BLEU-3	0.453	0.465	0.489	0.811
BLEU-4	0.229	0.237	0.258	0.604
BLEU-5	0.140	0.149	0.158	0.425

From **Table 2**, it can be observed that the BLEU scores for all models decrease as the n-gram length increases. This is due to the increased difficulty of matching the same sequences with longer n-grams, resulting in lower BLEU scores. However, even at BLEU-5 (the longest n-gram length), LeakGAN still achieves relatively high scores, indicating that it maintains a certain level of coherence and similarity to the target text during text generation.

LeakGAN consistently has the highest BLEU scores across all n-gram lengths, suggesting that it best preserves similarity to the target text, particularly at shorter n-gram lengths. As the n-gram length increases, BLEU scores for all models show a downward trend. However, the performance gap between different models may become more pronounced at longer n-grams. This highlights the importance of not only =considering overall performance when evaluating text generation models but also paying attention to how models perform across different n-gram lengths.

Table 3. Self-BLEU.

	SeqGAN	**MaliGAN**	**RankGAN**	**LeakGAN**
BLEU-2	0.877	0.868	0.893	0.963
BLEU-3	0.628	0.637	0.679	0.899
BLEU-4	0.402	0.417	0.463	0.809
BLEU-5	0.275	0.288	0.317	0.702

From **Table 3**, it can be observed that there are differences in Self-BLEU scores among the various models. However, overall, these scores are relatively low, indicating that all models exhibit a certain level of diversity in generating cyberbullying text. Lower Self-BLEU scores suggest that the models are capable of producing diverse and varied text, rather than merely repeating the same patterns or sentences.

LeakGAN has the potentially lowest Self-BLEU score, which indicates that it achieves the highest degree of diversity in generating cyberbullying text. This means that the texts generated by LeakGAN show considerable variation and are less prone to repetition.

5. Discussion and conclusion

LeakGAN significantly outperforms SeqGAN, MaliGAN, and RankGAN in generating Chinese cyberbullying text, achieving a BLEU2 score of 0.948, self-BLEU2 of 0.963, a low NLL value of 0.48, and the highest Embedding Similarity (EmbSim) at most time points. Its superior performance stems from its "leakage" mechanism, which grants the generator access to the discriminator's intermediate layer features. This strategy allows the generator to receive richer semantic feedback during training, resulting in outputs that more accurately reflect the structure, tone, and emotional content of real-world cyberbullying language. Consequently, LeakGAN generates text that is not only linguistically coherent but also psychologically realistic, aligning better with the affective patterns, aggression cues, and relational power imbalances observed in actual cyberbullying cases.

From a social psychological perspective, this capability is critical for studying how language reflects aggression, dominance, exclusion, and hostility behaviors commonly rooted in social learning theory, online disinhibition, and group conformity dynamics. By generating high-fidelity synthetic cyberbullying samples, LeakGAN enables the modeling of these psychological processes and allows researchers to simulate peer victimization, escalation of conflict, and bystander silence in digital environments. The synthetic data generated can thus support the development of psychologically grounded detection systems and deepen our understanding of how cyberbullying affects individual mental health, social reputation, and group cohesion.

In terms of online environmental context, LeakGAN's capacity to generate platform-specific language patterns allows for better simulation of behaviors across various ecosystems. For instance, WeChat-based cyberbullying often involves indirect exclusion (e.g., being removed from a group), while Weibo and Douyin expose victims to public shaming. LeakGAN's architecture enables it to adapt to the linguistic micro-environments of different platforms, thereby helping researchers study how interface features (e.g., anonymity, comment visibility) either facilitate or deter aggression. This aligns with Bronfenbrenner's Ecological Systems Theory, in which the macrosystem (social norms), exosystem (platform algorithms), and microsystem (peer interactions) jointly shape cyberbullying expression.

Methodologically, this study prioritizes psychological validity and cultural relevance in the evaluation of generated content. Emotional realism, intent coherence, and language appropriateness are assessed by native-speaking annotators familiar with cyberbullying discourse in China. These annotators can acutely capture the subtle emotional trend , nuanced shifts, and cultural connotations in the content.

For example, phrases like "人肉搜索" (doxxing) or "取关拉黑 "(unfollow and block) are culturally embedded cues that plain indicate digital aggression. Additionally, ethical safeguards such as screening for excessively graphic content and protecting annotators from emotional distress are implemented, ensuring alignment with ethical standards in psychological research.

While SeqGAN struggles with learning nuanced aggression due to limited feedback from the discriminator, and RankGAN or MaliGAN perform better on narrow semantic categories, LeakGAN is more effective in capturing the emotional diversity and implicit hostility characteristic of cyberbullying. However, its performance remains sensitive to hyperparameter settings, including the leakage strategy, discriminator complexity, and reward shaping. This sensitivity may affect its generalizability and consistency across different platforms or bullying types.

To address these challenges and further enhance the psychological robustness of the model, future work should explore hybrid architectures that integrate LeakGAN with pre-trained language models such

as GPT, BERT, or ERNIE (Baidu's Chinese-specific model). These models provide advanced contextual understanding, enabling LeakGAN to generate text that better reflects interpersonal nuance, contextual dependency, and emotional escalation key characteristics of real-world cyberbullying exchanges.

From a practical and applied perspective, the high-quality synthetic data generated by LeakGAN can be used to train detection models for schools, workplaces, and public online platforms. For example, in educational environments, it can be incorporated into AI-assisted early warning systems to flag potentially harmful peer interactions. In corporate settings, HR departments can use this data to develop digital misconduct detection tools that identify harassment in internal messaging platforms.

Furthermore, this approach enables cross-cultural analysis of cyberbullying, allowing researchers to compare how aggression manifests in different linguistic and cultural contexts. For example, comparison between Mandarin-speaking platforms and English-speaking ones can reveal how collectivist vs. individualist cultures influence bullying language and strategies.

Nonetheless, limitations remain regarding psychological generalizability. While synthetically generated text can replicate surface-level features and aggression styles, it may not fully capture the cognitive-emotional depth, personal history, or social motives behind real cyberbullying interactions.

Therefore, synthetic data should be triangulated with real-world data, interviews, or behavioral experiments to ensure ecological validity.

In conclusion, LeakGAN's architecture offers a powerful and socially relevant approach to generating synthetic Chinese cyberbullying text. Its outputs support not only technical advances in NLP, but also psychological insights into the structure, triggers, and impact of online aggression. By bridging machine learning with social psychology and digital environmental analysis, this work contributes to building more informed, ethical, and effective solutions for mitigating cyberbullying in digital ecosystems.

6. Acknowledgement

This work was partly supported by the Youth Project of Science and Technology Research Program of Chongqing Municipal Education Commission Open Fund Project Grant No. KJQN202301222, NO.KJQN202401235 and KJQN202401237.in part by Opening fund of Chongqing Engineering Research Center of Internet of Things and Intelligent Control Technology No.zhlv-20221031

Conflict of interest

The authors declare no conflict of interest

References

1. Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry 49, 376–385. doi:10.1111/j.1469-7610.2007.01846.x
2. Kaimba Frank.School intervention in peer cyberbullying: the case of high school students in Kafue District, Zambia [D]. Supervisor: MingHuo. Northeast Normal University,2023
3. Ding Haoran. Research on Cyberbullying Problem of Minors and Its Governance [D]. Supervisor: Xu Lei. Nanjing University of Posts and Telecommunications,2023.
4. Fu Brazier. An international comparative study of cyberbullying prevention programmes in primary schools[D]. Supervisor: QianSongling;Men Xinwei. Jilin University of Foreign Languages,2023.
5. 65.3% of surveyed youth said they or people around them have experienced cyber violence[EB/OL].(2023-6-20) https://baijiahao.baidu.com/s?id=1769178755106899787&wfr=spider&for=pc
6. Rao, M. E., and Rao, D. M. (2021). The mental health of high school students during the COVID-19 pandemic. Front. Educ. 6, 719539. doi:10.3389/feduc.2021.719539

7. Englander E. (2021). 3.5 social and mental health during the COVID- 19 ,pandemic. J. Am. Acad. Child Adolesc. Psychiat. 60, S147. doi: 10.1016/j.jaac.2021. 09.039
8. Paat, Y. F., and Markham, C. (2020). Digital crime, trauma, and abuse: Internet safety and cyber risks for adolescents and emerging adults in the 21st century. Soc. Work Ment. Health 19, 18–40. doi:10. 1080/15332985.2020.1845281
9. Kim, Y. J., Qian, L., and Aslam, M. S. (2020). Development of a personalized mobile mental health intervention for workplace cyberbullying among health practitioners: protocol for a mixed methods study. JMIR Res. Protoc. 9, e23112. doi: 10.2196/23112
10. Nochaiwong, S., Ruengorn, C., Thavorn, K., Hutton, B., Awiphan, R., Phosuya, C., et al. (2021). Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. Sci. Rep. 11, 1. doi:10.1038/s41598-021-89700-8
11. Kowalski, R. M., Toth, A., and Morgan, M. (2017). Bullying and cyberbullying in adulthood and the workplace. J. Soc. Psychol. 158,64–81. doi:10.1080/00224545.2017.1302402
12. Kowalski, R. M., Limber, S. P., and McCord, A. (2018). A developmental approachto cyberbullying: prevalence and protective factors. Aggress. Violent Behav. 45, 20–32.doi:10.1016/j.avb.2018.02.009
13. Goodfellow, I. J., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems.
14. Andreini, P., Ciano, G., Bonechi, S., Graziani, C., Lachi, V., Mecocci, A., ... & Bianchini, M. (2021). A two-stage GAN for high-resolution retinal image generation and segmentation. Electronics, 11(1), 60.
15. Huang, G., & Jafari, A. H. (2023). Enhanced balancing GAN: Minority-class image generation. Neural computing and applications,35(7), 5145-5154.
16. Tran, N. T., Tran, V. H., Nguyen, N. B., Nguyen, T. K., & Cheung, N. M. (2021). On data augmentation for GAN training. IEEE Transactions on Image Processing, 30, 1882-1897.
17. Liu, Y. (2021). Improved generative adversarial network and its application in image oil painting style transfer. Image and Vision Computing, 105, 104087.
18. Chen, Y., Zhang, H., Liu, L., Chen, X., Zhang, Q., Yang, K., ... & Xie, J. (2021). Research on image inpainting algorithm of improved GAN based on two-discriminations networks. Applied Intelligence, 51,3460-3474.
19. Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., ... & Guo, B. (2022). Styleswin: Transformer-based gan for high-resolution image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11304-11314).
20. Singh, N. K., & Raza, K. (2021). Medical image generation using generative adversarial networks: A review. Health informatics: A computational perspective in healthcare, 77-96.
21. Hossam, M., Le, T., Papasimeon, M., Huynh, V., & Phung, D. (2021). Text generation with deep variational GAN. arXiv preprint arXiv:2104.13488.
22. Chen, Z., Zhu, T., Xiong, P., Wang, C., & Ren, W. (2021). Privacy preservation for image data: a gan-based method. International Journal of Intelligent Systems, 36(4), 1668-1685.
23. Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017, February). Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).
24. Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., & Wang, J. (2018, April). Long text generation via adversarial training with leaked information. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
25. Ding, X., Wang, Y., Xu, Z., Welch, W. J., & Wang, Z. J. (2021, May). Ccgan: Continuous conditional generative adversarial networks for image generation. In International conference on learning representations.
26. Diao, S., Shen, X., Shum, K., Song, Y., & Zhang, T. (2021, August). TILGAN: transformer-based implicit latent GAN for diverse and coherent text generation. In Findings of the Association for Computational linguistics: ACL-IJCNLP 2021 (pp. 4844-4858).
27. Yang, Y., Dan, X., Qiu, X., & Gao, Z. (2020). FGGAN: Feature-guiding generative adversarial networks for text generation. IEEE Access, 8, 105217- 105225.
28. Deng Yang, Gao Kun, Liao Ning, Chen Yiran. Automatic script generation and optimisation technique based on generative adversarial network[J]. Digital Technology and Application,2024,42(02):232-234.
29. LI Bing,YANG Peng,SUN Yuankang,HU Zhongjian,YI Meng. Advances and Challenges in Artificial Intelligence Text Generation (in English)[J].Frontiers of Information Technology & Electronic Engineering,2024,25(01):64-84.
30. Xiong Lu,Pei Zhili,Jiang Mingyang & Bao Qiming. (2023). A text generation model based on improved generative adversarial network. Journal of Inner Mongolia University for Nationalities (Natural Science Edition) (02), 118-123. doi:10. 14045/j.cnki.15- 1220.2023.02.005.
31. Peng, P. F. & Zhou, L. R.. (2022). Adding reward to GRU adversarial network text generation model. Computers and Modernisation (07), 121- 126. doi:CNKI:SUN:JYXH.0.2022-07-019

32. Tingting Zhao, Yajing Song, Guixi Li, Lina Wang, Yarui Chen, Dehua Ren. A review of text generation research based on deep reinforcement learning[J]. Journal of Tianjin University of Science and Technology,2022,37(02):71-80.DOI:10. 13364/j.issn.1672-6510.20210146.
33. Q. Xue, X. F. Meng, F. Zhang, X. Y. Zhang, J. M. Zhu, Y. Zhu & D. D. Wang. (2022).HLMGAN: Hierarchical learning for multi-reward text generation adversarial networks. Journal of Yunnan University (Natural Science Edition)(01),64-72.
34. Y. Y. Kang, D. L. Peng, Z. Chen & C. C. Liu. (2019).ED-GAN: A legal text generation model based on improved generative adversarial networks. Small Microcomputer Systems (05), 1020-1025. doi:CNKI:SUN:XXWX.0.2019-05-021.
35. Zhou, W., Ge, T., Xu, K., Wei, F., & Zhou, M. (2020). Self-adversarial learning with comparative discrimination for text generation. arXiv preprint arXiv:2001.11691.
36. Mikolov T, Martin Karafiát, Burget L, et al.Recurrent neural network based language model[C]//Interspeech, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September.DBLP, 2015.DOI:10.1109/EIDWT.2013.25.
37. Yu L, Zhang W, Wang J, et al. Seqgan: Sequence generative adversarial nets with policy gradient[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
38. Che T, Li Y, Zhang R, et al. Maximum-likelihood augmented discrete generative adversarial networks[J]. arXiv preprintarXiv:1702.07983, 2017.
39. Lin K, Li D, He X, et al. Adversarial ranking for language generation[J]. Advances in neural information processing systems, 2017,30.
40. Guo J, Lu S, Cai H, et al. Long text generation via adversarial training with leaked information[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
41. Zhu Y, Lu S, Zheng L, et al. Texygen: A benchmarking platform for text generation models[C]//The 41st international ACM SIGIR conference on research & development in information retrieval. 2018:1097-1100.
42. Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.