frontiers in Physics

REVIEW
published: 29 November 2021
doi: 10.3389/fphy.2021.766540

Adversarial Machine Learning on Social Network: A Survey

Sensen Guo[1,2], *Xiaoyu Li*[1,2]* *and Zhiying Mu*[1,2]

[1]Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China, [2]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China

In recent years, machine learning technology has made great improvements in social networks applications such as social network recommendation systems, sentiment analysis, and text generation. However, it cannot be ignored that machine learning algorithms are vulnerable to adversarial examples, that is, adding perturbations that are imperceptible to the human eye to the original data can cause machine learning algorithms to make wrong outputs with high probability. This also restricts the widespread use of machine learning algorithms in real life. In this paper, we focus on adversarial machine learning algorithms on social networks in recent years from three aspects: sentiment analysis, recommendation system, and spam detection, We review some typical applications of machine learning algorithms and adversarial example generation and defense algorithms for machine learning algorithms in the above three aspects in recent years. besides, we also analyze the current research progress and prospects for the directions of future research.

OPEN ACCESS

Edited by:
Shudong Li,
Guangzhou University, China

Reviewed by:
Xinghua Li,
Xidian University, China
Gui-Quan Sun,
North University of China, China

***Correspondence:**
Xiaoyu Li
lixiaoyu@nwpu.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *29 August 2021*
Accepted: *14 October 2021*
Published: *29 November 2021*

Citation:
Guo S, Li X and Mu Z (2021) Adversarial Machine Learning on Social Network: A Survey. Front. Phys. 9:766540. doi: 10.3389/fphy.2021.766540

Keywords: social networks, adversarial examples, sentiment analysis, recommendation system, spam detection

1 INTRODUCTION

In recent years, with the rapid development of internet technology, social networks have played an increasingly important role in people's lives [1]. Among them, social networks such as Facebook, Twitter, and Instagram have shortened the distance between people and changed the way that people get information. For example, more and more people are willing to share new things happening around them with friends through social networks, and government agencies release the latest policy information to the public through social networks. With the rapid popularization of social networks, the role of social networks is not limited to providing people with a channel to communicate with friends. For example, users can be profiled according to its timelines, then the system can recommend friends, topics, information, and products that users may be interested in, which can greatly enrich people's leisure life. Filtering useless spam and robot accounts can not only reduce the time that users spend on browsing spam but also protect users from phishing website attacks. Besides, research on social network information dissemination [2, 3] can not only facilitate social network marketing but also effectively predict and control public opinion. The study of the interaction between disease and disease information on complex networks [4] has played an important role in understanding the dynamics of epidemic transmission and the interaction between information dissemination. Therefore, how to use social networks to achieve various functions has become a research hotspot in recent years.

With significant improvement in the performance of computers and the widespread application of GPUs, Machine learning (ML) especially Deep Learning (DL) has been widely used in various industries (such as automatic driving, computer vision, machine translation, recommendation

systems, cybersecurity, etc.). In terms of social networks, many scholars also use machine learning algorithms to implement functions such as friend or information recommendation, user interest analysis, and spam detection. However, it can't be ignored that machine learning algorithms are vulnerable to adversarial examples, that is, adding perturbations that are not perceptible to the human eye can mislead the classifier to output a completely different classification result. After the concept of adversarial examples was proposed, many studies have shown that no matter how the machine learning model is adjusted, it can always be successfully broken by new adversarial example generation methods. In recent years, the research on the generation and defense of adversarial examples has spread from the field of computer vision [5] to social networks, cybersecurity [6], natural language processing [5, 7], audio and video processing [8], graph data processing [5], etc. Therefore, the ability to effectively defend against adversarial examples has become a key factor of whether machine learning algorithms can be applied on a large scale.

In this paper, we focus on adversarial machine learning in the field of social networks, that is, adversarial example generation and defense technology in the field of social networks. Firstly, we reviewed the recent research progress of machine learning algorithms in social networks in terms of sentiment analysis, recommendation systems, and spam detection, and then we summarized the latest research on adversarial example generation and defense algorithms in recent years. Next we sorted out some research progress of adversarial example generation and defense algorithms in social networks. Finally, we summarized the advantages and disadvantages of existing algorithms, and prospects for its future research directions.

The rest of this paper is organized as follows. In **section 2**, the application of machine learning algorithms in social networks in recent years is reviewed. **Section 3** reviews the security issues faced by machine learning algorithms and the robust reinforcement strategies against different attacks. **Section 4** summarizes the attack and defense algorithms for machine learning in social networks. **Section 5** analyzes the problems of adversarial example generation and defense algorithms in the field of social networks, prospect the future research direction, and concludes this paper.

2 MACHINE LEARNING IN SOCIAL NETWORKS

While social network such as Twitter, Facebook, and Instagram facilitate people's communication, they also change people's lifestyles to a great extent. The application of machine learning in social networks also promotes the vigorous development of social networks to a large extent. The main applications of machine learning in social networks are as follows: sentiment analysis, recommendation system, spam detection, community detection [9], network immunization [10], user behavior analysis [11, 12], and other aspects. In this paper, we mainly review the application of machine learning in social networks from three aspects: sentiment analysis, recommendation system, and spam detection.

2.1 Sentiment Analysis

Millions of users have posted various opinions on social networks every day, involving daily life, news, entertainment, sports, and other aspects. The emotional of user's comment on different topics can be divided into positive, neutral, and negative categories. With the user's emotional tendency on different topics, we can learn the user's personality, value tendency, and other information. And then more targeted strategies can be used for specific users in activities such as topic dissemination and product promotion. Some researches of machine learning in sentiment analysis are shown in **Table 1**.

Wang et al. [26] introduced a multi-head attention-based LSTM model to perform aspect-level sentiment analysis, they carry out their experiment on the dataset of SemEval 2014 Task 4 [27], the results of the experiment show that their model is advantageously competitive in aspect-level classification. Based on this, Long et al. [15] introduced an improved method with bidirectional LSTM network and multi-head attention mechanism, they utilize the multi-head attention to learn the relevant information from a different representation subspace, and achieved 92.11% accuracy on comment dataset from Taobao.

To perform aspect-based sentiment analysis of Arabic Hotels' reviews, both SVM and deep RNN were used in Al-Smadi et al. [13]'s works, respectively. They evaluated their method on Arabic Hotels' reviews dataset. The results show that the performance of SVM is superior to the other deep RNN approach in the aspect category identification, opinion target expression extraction, and the sentiment polarity identification, but inferior to RNN approaches in the execution time required for training and testing.

By using the API provided by Twitter, Hitesh et al. [14] collected 18,000 tweets without retweets on the term Indian elections. Based on these data, they proposed a model that combined with word2vec and random forest model to perform sentiment analysis, and they used a Word2Vec feature selection model to extract features and then train a random forest model for sentiment analysis, and their final accuracy reaches 86.8%.

Djaballah et al. [16] proposed a method to detect content that incites terrorism on Twitter, they collected tweets related to terrorism in Arabic and manually classified these tweets in "tweets not inciting terrorism" and "tweets inciting terrorism". Based on Google's Word2vec method [17], they introduce a method of Word2vec by the weighted average to generate tweets feature vectors, then SVM and Random Forest classifiers were used for the prediction of sentiments. The experiments results show that their method can improve the prediction results of the Word2vec method [17] slightly.

Ho et al. [18] proposed a two-stage combinatorial model to perform sentiment analysis. In the first stage, they trained five machine learning algorithms: logistic regression, naive Bayes, multilayer perceptron, support vector machine and random forest with the same dataset. In the second stage, a combinatorial fusion is used to combine a subset of these five algorithms, and experiment results show that the combination of these algorithms can achieve better performance.

To capture precise sentiment expressions in aspect-based sentiment analysis for reasoning, Liu et al. [19] introduced a

TABLE 1 | Maching learning in sentiment analysis.

Authors	Introduced methods	Year	Datasets	Baseline
Al-Smadi et al. [13]	SVM and Deep RNN	2018	Arabic Hotels' reviews	—
Hitesh et al. [14]	Word2Vec & Random forest	2019	Twitter	BOW, TF-IDF
Long et al. [15]	BiLSTM-MHAT	2019	Taobao	CNN, BiLSTM, Attention-BiLSTM
Djaballah et al. [16]	SVM, Random Forest	2019	Twitter	Word2vec [17]
Ho et al. [18]	Combinatorial model	2019	Kaggle	LR, NB, RF, SVM, MLP
Liu et al. [19]	AS-Reasoner	2019	SemEval-2014, SemEval-2015	LSTM, TD-LSTM, TD-LSTM.etc
Yao et al. [20]	DSSA-H	2020	Twitter	SVM, RF
Umer et al. [21]	CNN-LSTM	2021	Twitter	CNN [22], LSTM [23]
Lv et al. [24]	CAMN	2021	SemEval-2014, Twitter	CEA, DAuM, TNet-AS,etc
Rawat et al. [25]	SMODT	2021	Twitter	KNN, SVM, DT, SMO

TABLE 2 | Maching learning in recommendation system.

Authors	Introduced methods	Year	Datasets	Baseline
Fan et al. [28]	GraphRec	2019	Epinions, Ciao	GC-MC [29], DeepSoR [30], NeuMF [31]
Gui et al. [32]	Cooperative Multi-Agent Approach	2019	Dataset Containing 50 Historical Tweets Per User	LSTM, Attention methods, Independent Q-Learning, Random sampling
Guo et al. [33]	GNN-SoR	2020	pinions [34], Yelp [35], Flixster [36]	SocialMF [37], TrustSVD [38], TrustMF [39], AutoRec [40]
Huang et al. [41]	MAGRM	2020	Meetup, MovieLens-1M	DPMF-CNN [42], AGR [43], AGREE [44]
Pan et al. [45]	CoDAE	2020	Epinions, Ciao	CDAE [46], TDAE [47]
Zheng et al. [48]	ITRA	2021	Delicious [49], FilmTrust [50], CiaoDVD [51]	CDAE [46], SAMN [52], CAVE [53]
Ni et al. [54]	RM-DRL	2021	Netflix, BookCrossing, Movielens-20M, Movielens-1M, HetRec 2011-Movielens	ConvMF [55], DRMF [56], GNN [57], AFM [58], RACMF [59], HRAM [60], DAINN [61]
Tahmasebi et al. [62]	SRDNet	2021	MovieTweetings, Open Movie Database	AutoRec [40], MRS-RBM [63], PP-CF [64], et

method named Attention-based Sentiment Reasoner (AS-Reasoner). In their model, an intra attention and a global attention mechanism was designed, respectively. The intra attention computes weights by capturing the sentiment similarity between any two words in a sentence, and the global attention computes weights by a global perspective. They carried out an experiment on various datasets, and the results show that the AS-Reasone is language-independent, and it also achieves state-of-the-art macro-F1 and accuracy for aspect-based sentiment analysis.

Umer et al. [21] proposed a deep learning model which is combined with CNN and LSTM network to perform sentiment analysis on Twitter. The CNN layer is used to learn the higher-level representation of sequences from original data and feed it to the LSTM layers. They carry out their experiment on three Twitter dataset which includes a women's e-commerce dataset, an airline sentiment dataset, and a hate speech dataset, and the accuracy on three datasets is 78.1, 82.0, and 92.0%, respectively, which is markedly superior to singly use of CNN [22] and LSTM [23].

2.2 Recommendation System

The social network recommendation system is an important part of the social network system. Recommendation systems such as friend recommendation, content recommendation, and advertising delivery greatly enrich people's social life while also create huge economic benefits. Recommending friends and article content that users may be interested in will extend the time users surf the social networks; Pushing advertising information to users reasonably and effectively can not only creating significant economic benefits but also facilitate users' lives. As shown in **Table 2**, with the rapid development of machine learning, many scholars have also carried out research on social network recommendation system based on machine learning.

Fan et al. [28] try to perform social recommendation with graph neural networks, and they introduced a model named GraphRec (Graph Neural Network Framework), which is composed of the user modeling, the item modeling, and the rating prediction. Both the user modeling and the item modeling used graph neural network and attention network to learn user latent factors ($\mathbf{h_i}$) and the learn item latent factors ($\mathbf{z_j}$) from the original data, respectively, the rating prediction concatenate the user latent factors and the item latent factors and feed into a multilayer perceptron neural network for rating prediction. They evaluated the GraphRec with two representative datasets Epinions and Ciao, and the results show that the GraphRec can outperform GC-MC (Graph Convolutional Matrix Completion) [29], DeepSoR (Deep Neural Network Model on Social Relations for Recommendation) [30], NeuMF (Neural Matrix Factorization) [31], and some other baseline algorithms.

Guo et al. [33] hold that the feature space of social recommendation is composed of user features and item feature, the user feature is composed of inherent preference and social influence, and the item feature include attribute contents, attribute correlations, and attribute concatenation. They introduced a framework named GNN-SoR (Graph

Neural Network-based Social Recommendation Framework) to exploit the correlations of item attributes. In their framework, two graphs neural network methods are used to encode the user feature space and the item feature space, respectively. Then, the encoded two spaces are regarded as two potential factors in the matrix factorization process to predict the unknown preference ratings. They conducted experiments on real-world datasets Epinions [34], Yelp [35] and Flixster [36] respectively, and the experimental results indicated that the perform of GNN-SoR is superior to four baselines algorithm such as: SocialMF (Matrix Factorization based Social Recommendation Networks) [37], TrustSVD [38], TrustMF [39], and AutoRec [40].

Huang et al. [41] introduced a model named MAGRM (Multiattention-based Group Recommendation Model) to perform group recommendation, and the MAGRM is consists of two multiattention based model: the VR-GF (vector representation for group features) and the PL-GI (preference learning for groups on items). The VR-GF is used for getting the deep semantic feature for each group. Based on VR-GF, the PL-GI is used for predicting groups' ratings on items, the experiment with two real-world dataset Meetup and MovieLens-1M, and the performance of MAGRM outperforms AGR [43], AGREE (Attentive Group Recommendation) [44] and other algorithms.

Pan et al. [45] introduced a model named CoDAE (Correlative Denoising Autoencoder) to perform top-k recommendation task, which learn user features by modeling user with truster, roles of rater, and trustee with three separate denoising autoencoder model. They carried out an experiment on Ciao and Epinions datasets, they found that their method is superior to CDAE (Collaborative Denoising Auto-Encoders) [46], TDAE [47], and some other baseline algorithms. Similar to [45], Zheng et al. [48]. proposed a model named ITRA (Implicit Trust Relation-Aware model) which is based on Variational Auto-Encoder to learn the hidden relationship between huge amounts of graph data. They evaluated their model on three dataset: Delicious [49], FilmTrust [50], and CiaoDVD [51], where the performance of ITRA was markedly superior to SAMN (Social Attention Memory Networ) [52], CVAE [53], and CDAE [46] in the top-n item recommendation task.

By capturing the semantic features of users and items effectively, Ni et al. [54] proposed a model named RM-DRL (Recommendation Model based on Deep Representation Learning). According to the authors, firstly, they used a CNN network to learn the semantic feature vector of the item from its primitive feature vectors. Next, they used an Attention-Integrated Gated Recurrent Unit to learn user semantic feature vector from a series of user features such as the user preference history, semantic feature vectors, primitive feature vector and so on. Finally, the users' preferences on the items were calculated with the semantic feature vectors of the items and the users. They conduct their experiments on five datasets, and the results show that the performance of RM-DRL is superior to ConvMF [55], AFM (Attentional Factorization Machines) [58], GNN [57], HRAM (Hybrid Recurrent Attention Machine) [60], etc.

2.3 Spam Detection

Social networking is one of the main channels for people to acquire information. However, the overwhelming spam and network phishing links also bring great troubles to people's work and life. Therefore, how to detect spam on social networks effectively is an important issue. As shown in **Table 3**, many scholars have proposed various methods to solve this problem in recent years.

Karakasli et al. [65] tried to detect spam users with machine learning algorithms. Firstly, they collect twitter user data with software named CRAWLER. Then, 21 features in total was extracted from the original Twitter data. Next, a dynamic feature selection method was used to reduce the model complexity. Finally, they used SVM and KNN algorithm to perform spam user detection, and the success detects rate for KNN was 87.6 and 82.9% for SVM.

Aiming at the problem of difficult spam detection caused by the short text and large semantic variability on social networks, by combining the convolutional neural network (CNN) with long short term memory neural network (LSTM), Jain et al. [66] introduced a deep learning spam detection architecture named Sequential Stacked CNN-LSTM (SSCL). Firstly, it uses the CNN network to extract feature sequences from original data, then it feed the feature sequences to the LSTM network, and finally the sigmoid function was used to classify the label as spam or none-spam. They evaluated the performance of SSCL on two dataset: SMS and Twitter, and its precision, accuracy, recall, and F1 score achieved 85.88, 99.01, 99.77, and 99.29%, respectively.

Zhao et al. [68] introduced a semi-supervised graph embedding model to detect spam bot for the directed social network, where they used the attention mechanism and graph neural network to detect spam bot based on the retweet relationship and the following relationship between users. They experimented with the Twitter 1KS-10KN dataset [69] which was collected on Twitter, compared with GCN, GraphSAGE, and GAT, their method achieved the best performance in Recall, Precision, and F1-score.

Focusing on the uneven distribution of spam data and non-spam data on Twitter, Zhang et al. [70] proposed an algorithm named I2RELM (Improved Incremental Fuzzy-kernel-regularized Extreme Learning Machine), which adopt fuzzy weights (each input data is provided with a weight s_i, which is in the interval of (0,1] and assigned by the ratio of spam users to non-spam users in the whole dataset) to improve the detection accuracy of the model on the non-uniformly distributed dataset. They evaluated their method with the data obtained from Twitter, and the performance of I2RELM on the accuracy, TRP, precision, and F-measure was superior to SVM, DT, RF, BP, RBF, ELM, and XG-Boost.

To perform spam detection for movie reviews, Gao et al. [71] proposed an attention mechanism based machine learning model named adCGAN. Firstly, they used SkipGram to extract word vectors from all reviews, and extended SIF algorithm [80] to generate sentence embedding. Then, they combined the encoded movie features and sentence vectors, and used attention driven generate adversarial network to perform review spam detection. They evaluated their method with the review data collected from

TABLE 3 | Machine learning in spam detection.

Authors	Introduced methods	Year	Datasets	Baseline
Karakasli et al. [65]	SVM and KNN	2019	Twitter	—
Jain et al. [66]	SSCL	2019	SMS and Twitter	KNN, NB, RF, SVM etc.
Tajalizadeh et al. [67]	INB-DenStream	2019	Twitter	DenStream, StreamKM++, CluStream
Zhao et al. [68]	Attention + GNN	2020	Twitter 1KS-10KN [69]	GCN, GraphSAGE, and GAT
Zhang et al. [70]	I2RELM	2020	Twitter	SVM, DT, RF, BP, RBF, ELM, XG-Boost
Gao et al. [71]	adCGAN	2020	Douban	MCSVM [72], VAE [73]
Zhao et al. [74]	Ensemble Learning	2020	[75]	CSDNN and WSNN [76]
Alom et al. [77]	Text-based & Combined classifier	2020	Twitter Social Honeypot, Twitter 1KS-10KN [69]	Blacklist-based Approach [78]
Neha et al. [79]	LSTM + Attention	2021	Twitter	Bi-LSTM, K Neighbor, Random forest, Decision tree, Naive Bayes

Douban, the accuracy of adCGAN achieved 87.3%, which was markedly superior to MCSVM [72], VAE [73], and some other baseline algorithms.

Aiming at the problem of class imbalances in the spam detection task, Zhao et al. [74] proposed an ensemble learning framework which based on heterogeneous stacking. Firstly, six different machine learning algorithms including SVM, CART, GNB (Gaussian Naive Bayes), KNN, RF, and LR were used to perform classification tasks separately. Then, feed the output of six machine learning algorithm to cost-sensitive learning based neural network to get the spam detect result. They experimented with the dataset collected by Chen et al. [75], and its performance was markedly superior to CSDNN and WSNN [76].

3 SECURITY IN MACHINE LEARNING

The concept of adversarial example was first introduced by Szegedy et al. [81], they found that the machine learning classifier would get completely different results by adding perturbation that hardly perceptible by the human eye to the original picture, Szegedy believes that the discontinuity of mapping between input and output caused by the highly nonlinear machine learning model is the main cause for the existence of adversarial examples. While Goodfellow et al. [82] and Luo et al. [83] believe that the machine learning model are vulnerable to adversarial examples is mainly due to its linear part, in the high-dimensional linear space, the superposition of multiple small perturbations in the network will cause a great change in the output. Glimer et al. [84] believe that adversarial examples are caused by the high dimensionality of the input data, while Ilyas et al. [85] believe that the adversarial example is not bugs but features, since the attributes of the dataset include robustness and non-robustness features, when we delete non-robust features from the original training set, we can obtain a robust model through training, the adversarial examples are generated due to its non-robust features, and have little relation with machine learning algorithms.

3.1 Attacks to Machine Learning Models

The generation process of adversarial examples is to mislead the target machine learning model by adding perturbation η that are imperceptible to the human eye on the original data, which can be expressed as [86]:

$$\min_{x^{adv}} J\left(f\left(x^{adv}\right), y^{adv}\right) \quad \text{s.t.} \quad \begin{cases} \|\eta\|_p \le \varepsilon, \\ f(x) = y, \\ y \ne y^{adv}, \end{cases} \tag{1}$$

where $J(\cdot)$ is the loss function, $f(\cdot)$ is the target machine learning model, x^{adv} is the adversarial example, η is the adversarial perturbation added to original data x, ε is a normal used to limit the size of η.

According to the degree of understanding to the target model, attacks to machine learning models can be divided into white-box attacks and black-box attacks. White-box attacker obtains all information such as the structure and parameters of the target model, on the contrary, the black-box attacker know nothing about the structural information of the target model, and can only query the output of the target model based on the input [87].

3.1.1 White-Box Attacks

Szegedy et al. [81] first introduced a white-box attack method named L-BFGS, which try to craft adversarial examples by defining the search for the smallest possible attack perturbation as an optimization problem, it can be expressed as:

$$\min_{x'} c\|\eta\| + J_\theta\left(x', l'\right) \quad \text{s.t.} \quad x' \in [0, 1] \tag{2}$$

where c is a constant, η is the perturbation, $J(\cdot)$ is the loss function.

Although L-BFGS has a high attack success rate, its computational complexity is expensive; Similar to Szegedy, based on optimization method, Carlini et al. [88] also proposed an adversarial example generate method named C&W. The research made by Carlini et al. showed that the algorithm can effectively attack most of the existing models [89, 90]; Combining C&W and Elastic Net, Chen et al. [91] introduced a method named EAD to craft adversarial examples, compared with C&W, the adversarial examples generated by EAD have stronger transferability.

To reduce the computation complexity of L-BFGS, Goodfellow et al. [82] introduced a method named FGSM

(Fast Gradient Sign Method), which is a single-step attack that adds perturbation along the direction of gradient, and the perturbation is calculated as $\eta = \varepsilon \text{sign}(\nabla_x J(\theta, x, y))$, where $J(\cdot)$ is the loss function, θ is the parameters of target model, and ε is the size of the perturbation; Based on FGSM, Kurakin et al. [92, 93]) introduced a method named BIM (Basic Iterative Method), which used an iterative method to generate adversarial examples, and they also used the real pictures to evaluate the effectiveness of BIM; Based on BIM, by limiting the size of the perturbation in each iteration strictly, Madry et al. [94] introduced a method named PGD (Projected Gradient Descent), the experiment result shows that the adversarial examples crafted by PGD have better transferability; Similar to Madry et al. [94], Dong et al. [95] introduced a method named MI-FGSM, which integrated the momentum term into the iterative process to craft adversarial examples, compare with BIM, the MI-FGSM can effectively escape from poor local maxima during the iterations.

In order to find the minimal perturbations that are sufficient to mislead the target machine learning model, based on the iterative linearization of the classifier, Moosavi-Dezfooli et al. [96] proposed the DeepFool algorithm, which helps the attacker to craft adversarial examples with minimal perturbations.

Without calculating the gradient of the target model, Baluja et al. [97] introduced a method named ATN (Adversarial Transformation Network) to perform white-box or black-box attacks by training an adversarial transformer network, which can transformer the input data into the target or untargeted adversarial examples. Similar to ATN, Xiao et al. [98] introduced advGan to craft adversarial examples, based on generative adversarial network, the generator of advGAN is used to generate the perturbation to the input data, and the discriminator is used to distinguish the original data from adversarial examples generated by the generator. Besides, Bai et al. [99] proposed AI-GAN to crafted adversarial examples. The above methods [97–99] only need to query the target model during the stage of model training stage, which is fast and efficient.

To find the strongest attack policy, Mao et al. [100] proposed Composite Adversarial Attacks (CAA). They adopted the NSGA-II genetic algorithm to find the best combination of attack algorithms from a candidate pool composed of 32 base attackers. The attack policy of CAA can be expressed as:

$$s: \mathcal{A}_N^s\left(\mathcal{A}_2^s\left(\mathcal{A}_1^s\left(x, \mathcal{F}; \epsilon_{s_1}, t_{s_1}\right), \mathcal{F}; \epsilon_{s_2}, t_{s_2}\right)\right) \ldots, \mathcal{F}; \epsilon_{s_N}, t_{s_N}\right) \quad (3)$$

where $\mathcal{A}_i(\cdot)$ is one of the attack algorithm in attack pool, ϵ_{s_i} and t_{s_i} is the hyperparameter of $\mathcal{A}_i(\cdot)$, $\mathcal{F}$ is the target model.

3.1.2 Black-Box Attacks

During the processes of black-box attack, the attacker know nothing about the target model, and the mainstream approach is based on gradient estimation and substitute model.

3.1.2.1 Based on Gradient Estimation

In this scenario, the attacker estimates the gradient information of the target model by feeding data into the target model and querying its output. Chen et al. [101] extended the C&W [88] algorithm and proposed Zeroth Order Optimization (ZOO) algorithm to perform black-box adversarial examples generation. Although the ZOO algorithm has a high success rate in generating adversarial examples, it requires a large amount of queries on the target model. To reduce the number of queries to the target model, Ilyas et al. [102] used the variant of NES algorithm [103] to estimate the gradient of the target model, which significantly reduces the query complexity to the target model. Tu et al. [104] proposed a framework named AutoZOOM, which adopts an adaptive random gradient estimation strategy and dimension reduction techniques to reduce the query count, compared with the ZOO [101], under the premise of achieving the same attack effect, AutoZOOM can significantly reduce the query complexity. Du et el [105]. also train a meta attacker mode to reduce the query count. Bai et al. [106] proposed the NP-attack algorithm, which also greatly reduces the query complexity by exploring the distribution of adversarial examples around benign inputs. Besides, Chen et al. [107] proposed the HopSkipJumpAttack algorithm, which applies binary information at the decision boundary to estimate gradient direction.

3.1.2.2 Based on Substitute Model

Based on the transferability of the adversarial examples, the attack usually trains a substitute model and uses the white-box attack algorithm to craft adversarial examples on the substitute model. Papernot et al. [108] first used substitute model to generate adversarial examples. Their research also shows that the attacker can perform black-box attack based on the transferability of adversarial examples, even if the structure of the substitute model is completely different from the target model. Zhou et al. [109] proposed a data-free substitute model train method (DaST) to train a substitute model for adversarial attack without any real data. By efficiently using the gradient of the substitute model, Ma et al. [110] proposed a highly query-efficient black-box adversarial attack model named SWITCH. Zhu et al. [111] used the PCIe bus to learn the information of machine learning models in the model-privatization deployments and proposed the Hermes Attack algorithm to fully reconstruct the target machine learning model. By focusing on the training strategy of the substitute model on the data distributed near the decision boundary, Wang et al. [112] improve the transferability of adversarial examples between the substitute model and the target model significantly. Based on meta-learning, Ma et al. [113] train a generalized substitute model named Simulator to mimic any unknown target model, which significantly reduces the query complexity to the target model.

3.2 Defense Against Adversarial Examples

The defense of adversarial examples is an important component of machine learning security. Many scholars have also proposed different adversarial example defense strategies in recent years. The strategies are divided into input data transformer, adversarial example detection, and model robust enhance.

3.2.1 Input Data Transformer

Since perturbation of adversarial examples are usually visually imperceptible, by compressing away these pixel manipulation,

Das et al. [114] introduced a defense framework based on JPEG compression. Cheng et al. [115] adopt a self-adaptive JPEG compression algorithm to defend against adversarial attacks of the video.

Based on generative adversarial network, Samangouei et al. [116] introduced the Defense-GAN to defend against adversarial attacks. By learning the distribution of unperturbed images, the Defense-GAN can generate the clean sample that approximates the perturbed images. Although the Defense-GAN could defend against most commonly attack strategies, its hyper-parameters is hard to train. Hwang et al. [117] also introduced a Purifying Variational Autoencoder (PuVAE) to purify adversarial examples, which is 130 times faster than Defense-GAN [116] in inference time. Besides, Lin et al. [118] introduced InvGAN to speed up Defense-GAN [116]. Zhang et al. [119] proposed an image reconstruction network based on residual blocks to reconstruct adversarial examples into clean images, in addition, adding random resize and random pad layer to the end of the reconstruction network is very effective in eliminating the perturbations introduced by iterative attacks.

3.2.2 Adversarial Example Detection

Just as the name implies, the adversarial example detection algorithms enhance the robustness of the machine learning model by filtering out adversarial examples in a large number of data sets, it detects adversarial examples mainly by learning the differences in characteristics and distribution between the adversarial examples and the normal data.

Among many works, Liu et al. [120] used the gradient amplitude to estimate the probability of modifications caused by adversarial attacks and applied steganalysis to detect adversarial examples. The experiment indicated that their method can accurately detect adversarial examples crafted by FGSM [82], BIM [92], DeepFool [96], and C&W [88]. Wang et al. [121] proposed a SmsNet to detect adversarial examples, which introduced a "SmsConnection" to extract statistical features and proposed a dynamic pruning strategy to prevent overfitting. The experiment indicated that the performance of SmsNet was superior to ESRAM (Enhanced Spatial Rich Model) [120] on detecting adversarial examples crafted by various attacks algorithms.

Noticing the sensitivity of adversarial examples to the fluctuations occurring at the highly-curved region of the decision boundary, Tian et al. [122] proposed Sensitivity Inconsistency Detector (SID) to detect adversarial examples, which achieved detection performance in detecting adversarial examples with small perturbation. Besides, based on the feature that adversarial examples are more sensitive to channel transformation operations than clean examples, Chen et al. [123] proposed a light-weighted adversarial examples detector based on adaptive channel transformation named ACT-Detector. The experiments show that the ADC-detector can defend against most adversarial attacks.

To lessen the dependence on prior knowledge of attacks algorithms, Sutanto et al. [124] proposed a Deep Image Prior (DIP) network to detect adversarial examples, they used a blurring network as the initial condition to train the DIP network only using normal noiseless images. In addition, it is applicable for real-time AI systems due to its faster detection speed for real images. Liang et al. [125] consider the perturbation crafted by adversarial attacks as a kind of noise, They use scalar quantization and smoothing spatial filter to implement an adaptive noise reduction for input images.

3.2.3 Model Robust Enhancement

Model robust enhancement mainly includes adversarial training and certified training. Adversarial training improves the model's immunity to adversarial examples by adding adversarial examples in its training set [126]. Certified training enhances model robustness by constraining the output space of each layer of the neural network under specific inputs during the training process.

3.2.3.1 Adversarial Training

Adversarial training is one of the effective methods to defend against the attacks from adversarial example, the process of adversarial training can be approximated by the following minimum-maximum optimization problem [94].

$$\min_{\theta} \mathbb{E}_{(x,y)\in\mathcal{D}}\left[\max_{\delta\in S} L\left(f_{\theta}(x+\delta), y\right)\right] \tag{4}$$

where D is the set of training data, $f(\cdot)$ is the target neural network, θ is the parameter of $f(\cdot)$, L is the loss function, and δ is the adversarial perturbation.

Szegedy et al. [81] first introduced the concept of adversarial training by training the neural network on the dataset composed of clean data and adversarial examples. Goodfellow et al. [82] also tried to enhance the robustness of the machine learning model by adding adversarial examples crafted by FGSM algorithm to the training set. Although it is more effective in defending against attacks from the FGSM algorithm, it is helpless against attacks from more aggressive algorithms such as C&W [88] and PGD [94]. Madry et al. [94] tried to enhance the robustness of neural networks from the lens of robust optimization, they used the saddle point formula to optimize the parameters of the network model, thereby reducing the loss of the model on adversarial examples. Although adversarial training with PGD [94] algorithm can significantly enhance the robustness of the model, the computational complexity is very expensive when training on large-scale datasets.

To reduce the computational complexity of adversarial training, Shafahi et al. [127] introduced a speed adversarial training method by updating both the model parameters and images perturbations for each update step, and its training speed is 3–30 times than that of PGD [94]. Zhang et al. [128] found that the adversarial perturbation was only coupled with the first layer of the neural network, based on this, they proposed an adversarial training algorithm named YOPO(You Only Propagate Once), by focusing adversary computation only on the input layer of the neural network, experiment indicated that the training efficiency of YOPO was 4–5 times than that of original PGD training [94]. Besides, the research of Wong et al. [129] shown that the combination of FGSM [82] and random initialization in

TABLE 4 | Attack to machine learning in social network.

Authors	Year	Method	Dataset	Baseline	Attack type		Aspect		
					Black-box	White-box	SA	SD	RS
Gao et al. [133]	2018	DeepWordBug	Enron spam emails, IMDB	Projected FGSM, Random + DeepWordBug Transformer	✓		✓	✓	
Vijayaraghavan et al. [134]	2019	AEG	IMBA, AG News	DeepWordBug [133], NMT-BT [135]	✓		✓		
Ren et al. [136]	2020	Lage Scale Adversarial Attack	IMBA, Rotten Tomatoes Movie Reviews	FGSM [82], DeepFool [96], Textbugger [137]		✓	✓		
Li et al. [138]	2020	BERT-Attack	AG News, IMDB, Yelp, FAKE, SNLI, MNLI	TextFooler [139], Genetic attack [140]	✓		✓	✓	
Nuo et al. [141]	2020	WordChange	Ctrip, JD.com	TF-IDF, TextRank	✓		✓	✓	
Li et al. [142]	2020	CLARE	Yelp, AG News, MNLI, QNLI	TextFooler [139], TextFooler + LM, BERTAttack	✓		✓		
Garg et al. [143]	2020	BAE	Amazon Yelp, IMDB, MR	TextFooler [139]	✓		✓		
Jin et al. [139]	2020	TextFooler	AG News, FAKER, MR, Yelp, IMDB	Textbugger [137]	✓		✓		
Maheshwary et al. [144]	2021	Hard Label Attack	AG News, Yahoo Answers, MR, IMDB, Yelp, SNLI, MNLI	TextFooler [139], PSO [145], AEG [134]	✓		✓		
Yang et al. [146]	2017	Co-visitation attack	YouTube, eBay, Amazon, Yelp	Popular-item-attack, Random-item-attack		✓			✓
Fang et al. [147]	2018	Graph Poisoning Attack	MovieLens-100K, Amazon Instant Video	Co-visitation attack [146]		✓			✓
Christakopoulou et al. [148]	2019	Oblivious Recommender System Attack	MovieLens-100K, MovieLens-1M	—		✓			✓
Sun et al. [149]	2020	NIPA	Cora, Citeseer, Pumbed	Random, Preferential, PGA		✓			✓
Song et al. [150]	2020	PoisonRec	Steam, MovieLens-1M and Amazon	Popular Attack, Random Attack, Middle Attack, Power Item Attack, ConsLOP	✓				✓
Chang et al. [151]	2020	GF-Attack	Cora, Citeseer, Pubmed	Random, Degree, RL-S2V, Aclass	✓				✓
Fang et al. [152]	2020	TNA	Yelp, Amazon, Digital Music	PGA [153], SGLD [153]		✓			✓
Lin et al. [154]	2020	AUSH	MovieLens-100K, Amazon, FilmTrus	Random, Segment, Bandwagon, DCGAN		✓			✓
Fan et al. [155]	2021	CopyAttack	MovieLens-10M & Flixster, MovieLens-20M & Netflix	RL-Generative, RandomAttack, TargetAttack	✓				✓
Zhan et al. [156]	2021	BBGA	Cora, Citesser, Cora-ML	DICE-BB, Random, Mettack, Aclass	✓				✓
Finkelshtein et al. [157]	2021	Single-Node Attack	Cora, CiteSeer, PubMed, Twitter-Hateful-Users	EdgeGrad	✓	✓			✓
Huang et al. [158]	2021	Poisoning Attack	Movielens-100K, Movielens-1M, Last.fm	Random, Bandwagon, MF		✓			✓
Wu et al. [159]	2021	TrialAttack	Movielens-100K, Movielens-1M, FilmTrust	Random, Average, PGA [153], TNA [152], AUSH [154]		✓			✓

adversarial training can significantly reduce training cost while achieving similar effects to original PDG training [94].

3.2.3.2 Certified Training

Gowal et al. [130] proved that the robustness to PGD [94] attack was not a true measure of robustness. They focus on research on formal verification, and they proposed a neural network verified training algorithm named IBP (Interval Bound Propagation). Although the IBP algorithm is not only computationally cheaper but also significantly reduce the verified error rate, its training process is unstable, especially in the initial stages of training, to enhance the stability of IBP. Zhang et al. [131] combined the IBP [130] algorithm and the tight linear relaxation algorithm named CROWN [132], and proposed a verified training algorithm named IBP-CROWM. The experiment results shown that both standard errors and verified errors of IBP-CROWN were outperforming than IBP [130].

4 SECURITY OF MACHINE LEARNING IN SOCIAL NETWORKS

Most of the existing researches related to adversarial examples are focusing on the field of image classification. However, the generation of adversarial examples in social networks needs to process data like text and graph, unlike images, text and graph are discrete in feature distribution, which makes it more difficult to craft adversarial examples with text or graph. In this section, we mainly review the researches of adversarial examples in sentiment analysis (SA), spam detection (SD), and recommendation systems (RS), as well as some researches on question and

TABLE 5 | Defend against to adversarial examples in social network.

Authors	Year	Method	Dataset	Baselines	Attacks	Aspect		
						SA	SD	RS
Pruthi et al. [160]	2019	Robust Word Recognition	SST, IMBA, Stanford Sentiment Treebank	data augmentation [154], adversarial training [82]	Swap, Drop, Keyboard, Add	✓	✓	
Jia et al. [161]	2019	Certified Robustness Rraining	IMDB, SNLI	Standard Training, Data Augmentation	Genetic attack [140]	✓		
Zhou et al. [162]	2019	DISP	SST-2, IMDB	Adversarial Data Augmentation (ADA), Adversarial Training (AT), Adversarial Training (AT)	Insertion, Deletion, Swap, Random, Embed	✓		
Si et al. [163]	2020	AMDA	SST-2, IMDB	Adversarial Data Augmentation (ADA)	TextFooler [139], PWWS [164]	✓		
Wang et al. [165]	2020	MUDE	Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993)	Enchant 3 spell checker, scRNN	Permutation, Insertion, Deletion, Substitution	✓	✓	
Shi et al. [166]	2020	Transformers Robustness Verify	Yelp, SST	IBP	—	✓		
Ye et al. [167]	2020	Safer	IMDB, Amazon	Certified Robustness Rraining [161], IBP	Genetic attack [140]	✓		
Mozes et al. [168]	2020	FGWS	SST-2, IMDB	DISP [162]	Genetic attack [140], PWWS [164]	✓		
Zeng et al. [169]	2021	RanMASK	AG News, SST-2	Safer [167]	TextFooler [139], Bert-Attack [138], DeepWordBug [133]	✓		
Wang et al. [170]	2021	TextFirewall	IMBA, Yelp	Adversarial Training, Spelling Check and Recovery (SCR), RSE	Deepwordbug [133], Genetic attack [140], PWWS [164]	✓	✓	
Karimi et al. [171]	2021	BAT	SemEval 2014 task 4, SemEval 2016 task 5	BERT [172]	Gradient attack [173]	✓		
Du et al. [174]	2019	FNCF	Movielens-100K, Movielens-1M	Distillation [175]	C&W [88]			✓
Tang et al. [176]	2019	AMR	Pinterest, Amazon	POP, MF-BPR, DUIF, VBPR	FGSM [82]			✓
Manotumruksa et al. [177]	2020	SAO	MovieLens, Beauty, Video, Foursquare, Brightkite, Yelp	MostPop, BPR, APR, SASRec, ASASRec	—			✓
Li et al. [178]	2020	SACRA	Yelp, Foursquare	WRMF, MMMF, BPRMF, CofiRank, CLiMF, USG, GeoMF, etc.	FGSM [82]			✓
Wang et al. [179]	2020	ATMBPR	Movielens-100K, Yelp	BPR, CDAE, MPR, AMF, MLP, NeuMF, LRML, JRL, etc.	FGSM [82]			✓
Shahrasbi et al. [180]	2020	Semi-Supervised Attack Detection	Instacart grocery	LSTM	—			✓
Wu et al. [181]	2021	APT	FilmTrust, MovieLens-100K, MovieLens-1M, Yelp	Adversarial Training (AT), PCMF	AUSH [154], TNA [152], PGA [153]			✓
Yi et al. [182]	2021	DAVE	Yelp, Digital Music, MovieLens-1M, MovieLens-100K, Pinterest	NeuMF, CDAE, CFGAN, APR, ACAE, AVB, VAEGAN, CVAE-GAN, RecVAE	AAE			✓

answer robot and neural machine translation. Since the data used in sentiment analysis and spam detection are both texts, they are similar in the generation and defense of examples, we will review them in one subsection, due to the relative lack of research on question and answering robots and neural machine translation, we will review them in one subsection. **Table 4** and **Table 5** has shown some algorithms in sentiment analysis, spam detection, and machine translation of adversarial example generation and defense studies in recent years.

4.1 Security in Sentiment Analysis and Spam Detection

4.1.1 Adversarial Attacks

The goal of the attack against sentiment analysis and spam detection system is to craft a text x' that is semantically similar to the original text x but can mislead the target classifier. It can be expressed as:

$$\min_{x'} S(x, x') \quad \text{s.t.} \quad F(x) \neq F(x') \tag{5}$$

where function $S(\cdot)$ is used to compute the semantic similarity between x and x', $F(\cdot)$ is the target model.

The adversarial example generation algorithm for texts is mainly by finding the keywords in the whole sentence that have a greater impact on the classification results and then adding perturbation to these keywords.

Gao et al. [133] proposed an effectively black-box text adversarial example generate method named DeepWordBug, and they introduced temporal tail score (TTS) and temporal score (TS) to evaluate the importance of words in the sentence. According to the authors, firstly, they calculate the TTS and TS by

querying the output of the target model after shielding some words, and then combine the value of TTS and TS to calculate the importance of every word in the whole sentence. It can be expressed as:

$$\begin{aligned} \mathrm{TS}(x_i) &= F(x_1, x_2, \ldots, x_{i-1}, x_i) - F(x_1, x_2, \ldots, x_{i-1}) \\ \mathrm{TTS}(x_i) &= F(x_i, x_{i+1}, x_{i+2}, \ldots, x_n) - F(x_{i+1}, x_{i+2}, \ldots, x_n) \\ Score(x_i) &= \mathrm{TS}(x_i) + \lambda(\mathrm{TTS})(x_i) \end{aligned} \quad (6)$$

where, $F(\cdot)$ is target machine learning model, and x_i is the i-th word in the sentence. Finally, they modify some characters in the keywords to generate text adversarial examples. Experiments have proved that although DeepWordBug can generate text adversarial examples with a high success rate, it will introduce grammatical errors and can be easily defended by grammar detection tools.

Vijayaraghavan et al. [134] proposed an Adversarial Examples Generator (AEG) based on reinforcement learning to craft none-target text adversarial examples, according to authors, they evaluated the effectiveness of the AEG algorithm on two target sentiment analysis convolutional neural networks: CNN-Word and CNN-Char, the experiment showed that the AEG model was able to fool the target sentiment analysis models with high success rates while preserving the semantics of the original text.

Li et al. [138] also proposed a word-level text adversarial examples generate algorithm named BERT-Attack. According to the authors, firstly, different from [133], they tried to find the vulnerable words in a sentence by masking each word and the query the target model for correct label. Then they used a pre-trained model Bert to replace vulnerable vocabulary with grammatically correct and semantically similar words. The process of calculating the vulnerability of each word can be expressed as:

$$I_{w_i} = F(S) - F(S_{\setminus w_i}) \quad (7)$$

where $F(\cdot)$ is target machine learning model, $S = [w_0, \ldots, w_i \ldots,]$ is the input text, and $S_{\setminus w_i} = [w_0, \ldots, w_{i-1}, [\mathrm{MASK}], w_{i+1}, \ldots,]$ is the text that replace the w_i with $[MASK]$.

Based on multiple modification strategies, Nuo et al. [141] proposed a black-box Chinese text adversarial example generate method named WordChange. Similar to the algorithm for calculating *TS* in **Eq. 6**, they search for keywords by gradually deleting a certain vocabulary in the sentence and then querying whether the output of the model has changed, and then applying "insert" and "swap" strategies on these keywords, thereby generating Chinese text adversarial examples that can fool the machine learning model.

Jin et al. [139] proposed a text adversarial examples generate method named TextFooler, which craft adversarial examples by finding the words that have the greatest impact on the output of the target model in the whole sentence and replacing it with words that share similar meanings with the original words. Although the replaced words in the adversarial examples generated by TextFooler are similar to the original words, it may not fit overall sentence semantics. To make the text adversarial examples more natural and free of grammatical errors, Similar to [138] Garg et al. [143] proposed a text adversarial example generation algorithm named BAE. According to the authors, firstly, they calculate the importance of each word in the text, and then choose a certain word and replace it with MASK or insert a MASK adjacent to it according to the importance of each word. Finally, they use the pre-trained language model BERT-MLM [183] to replace the mask with a word that fits the context. Similar to BAE [143], Li et al. [142] also introduced a pre-trained language model based text adversarial example generation algorithm named CLARE (ContextuaLized AdversaRial Example). Compared with BAE [143], CLARE has richer attack strategies and can generate text adversarial examples with varied lengths. Experiment showed that the text adversarial examples generated by BAE [143] and CLARE [142] were more fluent, natural and grammatical.

To attack text neural networks in hard label black-box setting where the attacker can only get the label output by the target model, Maheshwary et al. [144] utilized a Genetic Algorithm (GA) to craft text adversarial examples that share similar semantics with the original text. Experimental results show that on sentiment analysis tasks, their method can generate text adversarial examples with a higher success rate using smaller perturbation than algorithms such as TextFooler [139], PSO (Particle Swarm Optimization) [145], AEG [134], etc.

In addition, different from generating examples by replacing some words or characters in the text, Ren et al. [136] introduced a white-box text adversarial example generate model to generate text adversarial examples on large scale without inputting the original text, and their model is composed of a vanilla VAE-based generator and a series of discriminators. The generator is used to generate text adversarial examples, and the discriminators are used to make the adversarial examples of different labels crafted by *G* look more realistic. Their experiment showed that the proposed model could deceive the target neural network with high confidence.

4.1.2 Defense Against Adversarial Attacks

The current defense strategies for text adversarial examples are mainly divided into two aspects: adversarial example processing and model robustness enhancement. The adversarial example processing method mainly includes identifying the adversarial examples by detecting the misspellings and unknown words contained in the text and performing partial vocabulary replacement of the adversarial examples to convert them into clean text; The model robustness enhancement method enhances the model's defense ability against adversarial examples through methods such as adversarial training and formal verification.

4.1.2.1 Adversarial Example Processing

Adversarial example detection is an important way to detect adversarial examples in sentiment analysis and spam detection. Pruthi et al. [160] proposed RNN-based word recognizers to detect adversarial examples by detecting misspellings in the sentences, but it is hard to defend word-level attacks. By calculating the influence of words in texts, Wang et al. [170] proposed a general text adversarial examples detection algorithm named TextFirewall. They used it to defend the adversarial attacks from Deepwordbug [133], Genetic attack [140], and PWWS (Probability Weighted Word Saliency) [164], and the

average attack success rate decreased on Yelp and IMDB are 0.73 and 0.63%, respectively. Mozes et al. [168] also proposed adversarial example detection method named FGWS (Frequency-Guided Word Substitutions), and they tried to detect text adversarial example with the frequency properties of adversarial words and achieved a higher F1 score than DISP [162] in SST-2 and IMDB dataset. Besides, Wang et al. [165] also proposed a framework named MUDE (Multi-Level Dependencies) to detect adversarial word by taking advantage of both character and word level dependencies.

Zhou et al. [162] also introduced a framework named DISP (Discriminate Perturbations) to transform the text adversarial examples into clean text data. According to the authors, firstly, they identified the perturbed tokens in the input text with a perturbation discriminator, and then replaced the perturbed token with an embedding estimator. Finally, they recovered these tokens into a clean text with a KNN(k-nearest neighbors) algorithm. The experiment indicated that the DISP was outperforming the Adversarial Data Augmentation (ADA), Adversarial Training (AT), and Spelling Correction (SC) in terms of the efficiency and semantic integrity of the text adversarial examples.

4.1.2.2 Model Robustness Enhancement

As mentioned above, the algorithms to enhance the robustness of the NLP model mainly include adversarial training and formal verification. In terms of adversarial training, Si et al. [163] introduced a method named AMDA (Adversarial and Mixup Data Augmentation) to cover the larger proportion of the attack space during the process of adversarial training by crafting large amount of augmented training adversarial examples and feeding them to the machine learning model. They used AMDA to defend against attacks from PPWS [164] and TextFooler [139] on the data sets SST-2, AG News and IMBD, and achieved significant robustness gains in both Targeted Attack Evaluation (TAE) and Static Attack Evaluation (SAE). For large pre-training model BERT, Karimi et al. [171] introduced a method named BAT to fine-tuned the BERT model by using normal and adversarial text at the same time to obtain a model with better robustness and generalization ability. The experiment indicated that the BERT model trained with BAT was more robust than the traditional BERT model in aspect-based sentiment analysis task.

In terms of formal verification, Jia et al. [161] proposed certified robustness training by using interval bound propagation to minimize the upper bound on the worst-case loss. Facts have proved that this method can effectively resist word substitution attacks from Genetic attack [140]. Shi et al. [166] proposed a transformers robustness verify method to verify the robustness transformers network, compared with the interval boundary propagation algorithm, their method could achieve much tighter certified robustness bounds. Ye et al. [167] proposed a structure-free certified robustness framework named SAFER, which only needs to query the output of the target model when verifying its robustness, so it can be applied to neural network models with any structure, but it is only suitable for word substitutions attacks. Zeng et al. [169] proposed a smoothing-based certified defense method named RanMASK, it could defend against both defense method against both the character and word substitution-based attacks.

4.2 Security in Social Recommendation System

4.2.1 Adversarial Attacks

The poisoning attack affects the recommendation list of the target recommendation system by feeding fake users into the recommendation system, which has occupies the dominant position in adversarial attacks against machine learning-based recommendation systems.

Yang [146] performed promotion and demotion poisoning attacks by taking attacks as constrained linear optimization problems, and they verified their method on real social network recommendation systems, such as YouTube, eBay, Amazon, Yelp, etc., and achieved a high success attack rate. Similar to Yang [146], Fang et al. [147] also formulates the poisoning attacks to graph-based recommendation system as an optimization problem, and performs poison attacks by solving these optimization problems. Christakopoulou et al. [148] proposed a two-step white-box poisoning attack framework to fool the machine learning-based recommendation system. Firstly, they utilize a GAN network to generate faker users, and then craft the profiles of fake users with projected gradient method. Fang et al. [152] performed attacks to matrix factorization based social recommendation system by optimizing the ratings of a fake user with a subset of influential users. Huang et al. [158] also tried to poison the deep learning based recommendation system by maximizing the hit rate of a certain item appearance in the top-n recommendation list predicted by target recommendation system.

To effectively generate fake user profile with strong attack power for poisoning attacks, Wu et al. [159] introduced a flexible poisoning framework named TrialAttack, the TrialAttack is based on GAN network and consists of three parts: generator G, influence module I, and discriminator D, the generator is used to generate fake user profile that is close to the real user and has attack influence, the influence model is used to guide the generator to generate fake users profile with greater influence, and the discriminator is used to distinguish the faker profiles generated by the generator from the real.

The above attack methods are all white-box-based attack algorithms, that is, the attacker needs to fully understand the parameter information of the target model, but this is unrealistic to the recommendation system in the real social network. In terms of black-box attacks, Fan et al. [155] introduced a framework named CopyAttack to perform a black-box adversarial attack to recommendation system in social network, they used reinforcement learning algorithms to select users from the original domain and inject them into the target domain to improve the hit rate of the target item in the top-n recommendation list.

Song et al. [150] proposed an adaptive data poisoning framework named PoisonRec, it leverages reinforcement learning to inject false user data into the recommendation

system, which can automatically learn effective attack strategies for various recommendation systems with very limited knowledge.

To attack the graph embedding models with limited knowledge, Chang et al. [151] introduced an adversarial attacker framework named GF-Attack, which formulated the graph neural network as a general graph signal processing with corresponding graph filters, and then attacked the graph filters through the feature matrix and adjacency matrix. To minimize the modification of the original graph data in the attack, Finkelshtein et al. [157] introduced a single-node attack to perform adversarial attack to graph neural networks, which could fool the target model by only modifying a single arbitrary node in the graph.

4.2.2 Defense Against Adversarial Attacks

The current defense algorithms for recommendation systems are mainly divided into two aspects: model robustness enhancement and abnormal data detection. Among them, model robustness enhancement is based on adversarial training, and abnormal data detection improves the robustness of the recommendation systems by recognizing pollution data.

In terms of adversarial training, Tang et al. [176] proposed an adversarial training method named AMR (Adversarial Multimedia Recommendation) to defend against adversarial attack. According to the authors, the process of adversarial training could be interpreted as playing a minimax game. On the one hand, continuously generate perturbations that can maximize the loss function of target model. On the other hand, continuously optimize the parameters of target model to identify these perturbations.

By combining knowledge distillation with adversarial training, Du et al. [174] produced a more robust collaborative filtering model based on neural network to defend against adversarial attacks. The experiments indicated that their model can effectively enhance the robustness of the recommendation system under the attack of the C&W [88] algorithm.

Manotumruksa et al. [177] also proposed a recommendation system robust enhancement framework named SAO (Sequential-based Adversarial Optimization) to enhance the robustness of the recommendation system by generating a sequence of adversarial perturbations and adding it into the training set during the training process.

Li et al. [178] introduced a framework named SACRA (Self-Attentive prospective Customer Recommendation Framework) to perform prospective customer recommendation. Similar to Manotumruksa [177], the SACRA enhances its robust by adding adversarial perturbation into the training set dynamically to make the recommend system immune to these perturbations.

Wu et al. [181] used the influence function proposed by Koh et al. [184] to craft fake users, and then injected these fake users into the training set to enhance the robustness of the recommendation system. They named their method as adversarial poisoning training (APT), they used five poisoning attack algorithms to evaluate the effectiveness of the APT. The experiment indicated that APT can enhance the robustness of the recommendation system effectively.

By combining the advantages of adversarial training and VAE (Variational Auto-Encoder), Yi et al. [182] proposed a robust recommendation model named DAVE (Dual Adversarial Variational Embedding), which is composed of User Adversarial Embedding (UserAVE), User Adversarial Embedding (ItemAVE) and Neural Collaborative Filtering Network, UserAVE and ItemAVE generate user and item embedding according to user interaction vector and item interaction vector, respectively. Then the user and item embedding are fed into the Collaborative Filtering Network to predict and recommend results. During the training process of the DAVE, it reduces the impact of adversarial perturbation by adaptively generating a unique embedding distribution for each user and item.

In terms of abnormal data detection, Shahrasbi et al. [180] proposed a GAN-based pollution data detection method. According to the authors, firstly, they convert the clean user session data to embedding sequences with a Doc2Vec language model. Then, during the training process of GAN, the generator is trained to learn the distribution of real embedding sequences, and the discriminator is trained to learn the distinguish the real embedding sequences and the sequence generated by the generator. Based on this, when the training of GAN network is completed, the pollution data can be identified from the whole dataset.

4.3 Security in Other Aspects of Social Networks

In this subsection, we mainly review some research on adversarial examples from the aspects of question and answer robot and neural machine translation.

4.3.1 Question and Answer Robot

Xue et al. [185] introduced a text adversarial example craft method named DPAGE (Dependency Parse-based Adversarial Examples Generation) to perform black-box adversarial attack to Q&A robots. They extract the keywords of the sentence based on the dependency relation of the sentences and then replace these keywords with the adversarial word that are similar to these keywords to craft adversarial questions. They evaluated the performance of DPAGE with two Q&A robots: DrQA and Google Assistant, and the results shown that the adversarial examples crafted by DPAGE cannot affect both the correct answer and the top-k candidate answers output by the Q&A robot. Similar to [185] Deng et al. [186] proposed a method named APE (Attention weight Probability Estimation) to extract keywords from the dialogue and fool the target Q&A system by replaced these keywords with synonyms. The experiment results show that their method can attack the Q&A system with a high success rate.

4.3.2 Neural Machine Translation

The NMT model is also vulnerable to attacks from adversarial examples. Ebrahimi et al. [187] proposed a white-box gradient-based optimization text adversarial example generation method to perform targeted adversarial attacks to NMT models. The

experiment results have shown that their method can attack the target NMT model with a high success rate, and the robustness of the model can improve significantly after robust training. Besides, in the study of poison attacks, Wang et al. [188] can successfully implement the poison attack by injecting only 0.02% of the total data into the data set.

To enhance the robustness of the NMT model, Cheng et al. [189] craft text adversarial examples with a white-box gradient-based method and then used it to enhance the robustness of the model. Experiments on English-German and Chinese-English translation tasks have shown that their method can significantly improve the robustness and performance of the NMT model. In another study by Cheng et al. [190], they also try to enhance the robustness of the NMT models by augmenting the training data with an adversarial augmentation technique.

5 DISCUSSION AND CONCLUSION

5.1 Discussion

Although the generation and defense algorithms of adversarial examples have made great achievements on unstructured data in social networks, there are still many key issues that have not been resolved.

5.1.1 Constraints for Attacks on Real Systems

Many adversarial example generation algorithms in social networks do not consider the restrictions on attacks on real systems. In terms of text adversarial generation, many studies [133, 138, 139, 141] try to get the keywords in the sentence by frequently querying the target model, however, the action of the frequent query is easy to be detected and defended when the attack is performed on the real system. In terms of adversarial example generation in the recommendation system, the attacker poisons the recommendation system by modifying the edge and attribute information of some nodes in the social network graph [146, 147, 150, 151], however, in the real social network, the node that the attacker chooses to modify may be a node that is not controlled by the attacker. Therefore, in the subsequent research on adversarial example generation algorithms in the field of social networks, more consideration should be given to the limitations in real scenarios.

5.1.2 The Security of Social Network Data

To adapt to the complex and changeable user structure on social networks, the rapid change and short timeliness of cyber language, the AI models applied to social networks need to frequently fine-tune its parameters based on real data from social networks. Therefore, poisoning attacks must be effectively avoided during the process of online training. Since adversarial examples are usually difficult to find visually, on the one hand, there is currently little research on adversarial examples detection for unstructured data such as graphs. On the other hand, with the continuous evolution of attack methods, the existing data enhancement and cleaning technologies cannot effectively detect malicious data in all data. Therefore, how to accurately detect poisoning data in social networks will become a focus of future research.

5.1.3 Robustness Evaluation of Social Network Models

Due to the poor interpretability of machine learning algorithms, it is difficult to analyze and prove the robustness of machine learning mathematically. Therefore, the current robustness evaluation of machine learning algorithms mainly depends on the defensive ability of specific adversarial attacks, however, the robustness conclusions obtained by this method are difficult to apply to the latest attack algorithms. In the field of computer vision, some researches [130, 131] have tried to use formal verification to analyze the robustness of machine learning algorithms. In terms of social networks, some researches [161, 166, 167, 169] also try to use formal verification algorithms to analyze the robustness of text classification machine learning models, but it is great affected by the model structure and data types, in terms of recommendation systems, the research on the robustness analysis for machine learning algorithm is still blank. Therefore, the robustness analysis for machine learning algorithm will also be a research focus in social networks.

5.2 Conclusion

Although machine learning algorithms have made significant developments in many fields, it cannot be ignored that machine learning algorithms are vulnerable to attacks from adversarial examples. That is, adding perturbations that are not detectable by the human eye to the original data may cause the machine learning algorithm to make a completely different output with a high probability. In this paper, we review the application of machine learning algorithms in the field of social networks from aspects of sentiment analysis, recommendation systems, and spam detection, as well as the research progress of the generation of adversarial examples and defense algorithms in social networks.

Although the data processed by machine learning models that are used in social networks is usually unstructured data such as text or graphs, the adversarial example generation algorithm for images in the field of computer vision is also applicable to unstructured data such as text and graphs after extension. Therefore, how to use machine learning algorithms to implement various functions in social networks while ensuring the robustness of the algorithm itself is one of the hotspots for studying. Besides, to improve the robustness of machine learning algorithms in the field of the social network, in terms of adversarial example generation, more focus should be put on the adversarial example generation algorithms that can be applied to real online social network machine learning models, so as to enhance the robustness of online machine learning models. In terms of adversarial example defense, while strengthening robustness against specific attacks, more research on active defense algorithms such as certified training should also be carried out to defend against adversarial attacks.

AUTHOR CONTRIBUTIONS

SG: Investigation, Analysis of papers, Drafting the manuscript, Review; XL: Review, Editing; ZM: Review, Editing.

FUNDING

This work was supported by National Key R&D Program of China (Grant No. 2020AAA0107700), Shenzhen Fundamental Research Program (Grant No. 20210317191843003), the Shaanxi Provincial Key R&D Program (Grant No. 2021ZDLGY05-01), the Natural Science Basic Research Plan in Shaanxi Province of China (Grant No. 2020JQ-214).

REFERENCES

1. Clark JL, Algoe SB, Green MC. Social Network Sites and Well-Being: the Role of Social Connection. *Curr Dir Psychol Sci* (2018) 27(1):32–7. doi:10.1177/0963721417730833
2. Liu C, Zhou N, Zhan X-X, Sun G-Q, Zhang Z-K. Markov-based Solution for Information Diffusion on Adaptive Social Networks. *Appl Maths Comput* (2020) 380:125286. doi:10.1016/j.amc.2020.125286
3. Li L, Zhang J, Liu C, Zhang H-T, Wang Y, Wang Z. Analysis of Transmission Dynamics for Zika Virus on Networks. *Appl Maths Comput* (2019) 347: 566–77. doi:10.1016/j.amc.2018.11.042
4. Zhan X-X, Liu C, Zhou G, Zhang Z-K, Sun G-Q, Zhu JJH, et al. Coupling Dynamics of Epidemic Spreading and Information Diffusion on Complex Networks. *Appl Maths Comput* (2018) 332:437–48. doi:10.1016/j.amc.2018.03.050
5. Han X, Yao M, Liu H-C, Deb D, Liu H, Tang J-L, et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *Int J Automation Comput* (2020) 17(2):151–78. doi:10.1007/s11633-019-1211-x
6. Guo S, Zhao J, Li X, Duan J, Mu D, Xiao J. A Black-Box Attack Method against Machine-Learning-Based Anomaly Network Flow Detection Models. *Security Commun Networks* (2021) 2021:1–13. doi:10.1155/2021/5578335
7. Zhang WE, QuanSheng Z, F Alhazmi AA, Li C. *Generating Textual Adversarial Examples for Deep Learning Models: A Survey* (2019). arXiv preprint arXiv:1901.06796.
8. Yao Q, Carlini N, Cottrell G, Goodfellow I, Raffel C. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In: 36th International Conference on Machine Learning, ICML 2019. Long Beach, CA, United states: PMLR (2019). p. 5231–40.
9. Li S, Jiang L, Wu X, Han W, Zhao D, Wang Z. A Weighted Network Community Detection Algorithm Based on Deep Learning. *Appl Maths Comput* (2021) 401:126012. doi:10.1016/j.amc.2021.126012
10. Li S, Zhao D, Wu X, Tian Z, Li A, Wang Z. Functional Immunization of Networks Based on Message Passing. *Appl Maths Comput* (2020) 366:124728. doi:10.1016/j.amc.2019.124728
11. Han W, Tian Z, Huang Z, Li S, Jia Y. Topic Representation Model Based on Microblogging Behavior Analysis. *World Wide Web* (2020) 23(6):3083–97. doi:10.1007/s11280-020-00822-x
12. Nie Y, Jia Y, Li S, Zhu X, Li A, Zhou B. Identifying Users across Social Networks Based on Dynamic Core Interests. *Neurocomputing* (2016) 210:107–15.
13. Al-Smadi M, Qawasmeh O, Al-Ayyoub M, Jararweh Y, Gupta B. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *J Comput Sci* (2018) 27: 386–93. doi:10.1016/j.jocs.2017.11.006
14. Hitesh MSR, Vaibhav V, Kalki YJA, Kamtam SH, Kumari S. Real-time Sentiment Analysis of 2019 Election Tweets Using Word2vec and Random forest Model. In: 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT). IEEE (2019). p. 146–51. doi:10.1109/icct46177.2019.8969049
15. Long F, Zhou K, Ou W. Sentiment Analysis of Text Based on Bidirectional Lstm with Multi-Head Attention. *IEEE Access* (2019) 7:141960–9. doi:10.1109/access.2019.2942614
16. Djaballah KA, Boukhalfa K, Omar B. Sentiment Analysis of Twitter Messages Using Word2vec by Weighted Average. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE (2019). p. 223–8. doi:10.1109/snams.2019.8931827
17. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: *Advances in Neural Information Processing Systems* (2013), Lake Tahoe, NV, United states. p. 3111–9.
18. Ho J, Ondusko D, Roy B, Hsu DF. Sentiment Analysis on Tweets Using Machine Learning and Combinatorial Fusion. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE (2019). p. 1066–71. doi:10.1109/dasc/picom/cbdcom/cyberscitech.2019.00191
19. Liu N, Shen B, Zhang Z, Zhang Z, Mi K. Attention-based Sentiment Reasoner for Aspect-Based Sentiment Analysis. *Human-centric Comput Inf Sci* (2019) 9(1):1–17. doi:10.1186/s13673-019-0196-3
20. Yao F, Wang Y. Domain-specific Sentiment Analysis for Tweets during Hurricanes (Dssa-h): A Domain-Adversarial Neural-Network-Based Approach. *Comput Environ Urban Syst* (2020) 83:101522. doi:10.1016/j.compenvurbsys.2020.101522
21. Umer M, Ashraf I, Mehmood A, Kumari S, Ullah S, Sang Choi G. Sentiment Analysis of Tweets Using a Unified Convolutional Neural Network-long Short-term Memory Network Model. *Comput Intelligence* (2021) 37(1): 409–34. doi:10.1111/coin.12415
22. Conneau A, Schwenk H, Barrault L, Lecun Y. Very Deep Convolutional Networks for Text Classification. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017-Proceedings of Conference. Valencia, Spain (2016). p. 1107–16. doi:10.18653/v1/e17-1104
23. Cliche M. Bb_twtr at Semeval-2017 Task 4: Twitter Sentiment Analysis with Cnns and Lstms. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017. Vancouver, Canada: Association for Computational Linguistics (2017). p. 573–80. doi:10.18653/v1/S17-2094
24. Lv Y, Wei F, Cao L, Peng S, Niu J, Yu S, et al. Aspect-level Sentiment Analysis Using Context and Aspect Memory Network. *Neurocomputing* (2021) 428: 195–205. doi:10.1016/j.neucom.2020.11.049
25. Rawat R, Mahor V, Chirgaiya S, Shaw RN, Ghosh A. Sentiment Analysis at Online Social Network for Cyber-Malicious post Reviews Using Machine Learning Techniques. *Computationally Intell Syst their Appl* (2021) 950: 113–30. doi:10.1007/978-981-16-0407-2_9
26. Wang Y, Huang M, Zhu X, Zhao L. Attention-based Lstm for Aspect-Level Sentiment Classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing (2016). p. 606–15. doi:10.18653/v1/d16-1058
27. Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Al-Smadi M, et al. Semeval-2016 Task 5: Aspect Based Sentiment Analysis. In: International workshop on semantic evaluation (2016). p. 19–30. doi:10.18653/v1/s16-1002
28. Fan W, Yao M, Li Q, Yuan H, Zhao E, Tang J, et al. Graph Neural Networks for Social Recommendation. In: The World Wide Web Conference (2019). p. 417–26. doi:10.1145/3308558.3313488
29. van den Berg R, Kipf TN, Welling M. *Graph Convolutional Matrix Completion* (2017). arXiv preprint arXiv:1706.02263.
30. Fan W, Li Q, Cheng M. Deep Modeling of Social Relations for Recommendation. In: *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).
31. He X, Liao L, Zhang H, Nie L, Hu X, Chua T-S. Neural Collaborative Filtering. In: Proceedings of the 26th international conference on world wide web (2017). p. 173–82. doi:10.1145/3038912.3052569
32. Gui T, Liu P, Zhang Q, Zhu L, Peng M, Zhou Y, et al. Mention Recommendation in Twitter with Cooperative Multi-Agent Reinforcement Learning. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019). p. 535–44. doi:10.1145/3331184.3331237
33. Guo Z, Wang H. A Deep Graph Neural Network-Based Mechanism for Social Recommendations. *IEEE Trans Ind Inform* (2020) 17(4):2776–83. doi:10.1109/tii.2020.2986316
34. Massa P, Avesani P. Controversial Users Demand Local Trust Metrics: An Experimental Study on Epinions. Com Community. *AAAI* (2005) 5: 121–6. doi:10.5555/1619332.1619354
35. Hegde S, Satyappanavar S, Setty S. Restaurant Setup Business Analysis Using Yelp Dataset. In: 2017 International Conference on Advances in Computing,

Communications and Informatics (ICACCI). IEEE (2017). p. 2342–8. doi:10.1109/icacci.2017.8126196
36. Yang X, Steck H, Guo Y, Liu Y. On Top-K Recommendation Using Social Networks. In: Proceedings of the sixth ACM conference on Recommender systems (2012). p. 67–74. doi:10.1145/2365952.2365969
37. Jamali M, Ester M. A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks. In: Proceedings of the fourth ACM conference on Recommender systems (2010). p. 135–42. doi:10.1145/1864708.1864736
38. Guo G, Zhang J, Yorke-Smith N. Trustsvd: Collaborative Filtering with Both the Explicit and Implicit Influence of User Trust and of Item Ratings. In: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29 (2015).
39. Yang B, Lei Y, Liu J, Li W. Social Collaborative Filtering by Trust. *IEEE Trans Pattern Anal Mach Intell* (2016) 39(8):1633–47. doi:10.1109/TPAMI.2016.2605085
40. Sedhain S, Menon AK, Scott S, Xie L. Autorec: Autoencoders Meet Collaborative Filtering. In: Proceedings of the 24th international conference on World Wide Web (2015). p. 111–2.
41. Huang Z, Xu X, Zhu H, Zhou M. An Efficient Group Recommendation Model with Multiattention-Based Neural Networks. *IEEE Trans Neural Netw Learn Syst*. (2020) 31(11):4461–74. doi:10.1109/tnnls.2019.2955567
42. Wang H, Dong M. Latent Group Recommendation Based on Dynamic Probabilistic Matrix Factorization Model Integrated with Cnn. *J Comput Res Dev* (2017) 54(8):1853. doi:10.7544/issn1000-1239.2017.20170344
43. Tran LV, Pham T-AN, Tay Y, Liu Y, Gao C, Li X. Interact and Decide: Medley of Sub-attention Networks for Effective Group Recommendation. In: Proceedings of the 42nd International ACM SIGIR conference on research and development in information retrieval (2019). p. 255–64.
44. Cao D, He X, Miao L, An Y, Yang C, Hong R. Attentive Group Recommendation. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (2018). p. 645–54. doi:10.1145/3209978.3209998
45. Pan Y, He F, Yu H. A Correlative Denoising Autoencoder to Model Social Influence for Top-N Recommender System. *Front Comput Sci* (2020) 14(3):1–13. doi:10.1007/s11704-019-8123-3
46. Wu Y, DuBois C, Zheng AX, Ester M. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In: Proceedings of the ninth ACM international conference on web search and data mining (2016). p. 153–62. doi:10.1145/2835776.2835837
47. Wang M, Wu Z, Sun X, Feng G, Zhang B. Trust-aware Collaborative Filtering with a Denoising Autoencoder. *Neural Process Lett* (2019) 49(2):835–49. doi:10.1007/s11063-018-9831-7
48. Zheng Q, Liu G, Liu A, Li Z, Zheng K, Zhao L, et al. Implicit Relation-Aware Social Recommendation with Variational Auto-Encoder. *World Wide Web*. Springer (2021). p. 1–16.
49. Cantador I, Peter B, Kuflik T. Second Workshop on Information Heterogeneity and Fusion in Recommender Systems (Hetrec2011). In: Proceedings of the fifth ACM conference on Recommender systems (2011). p. 387–8. doi:10.1145/2043932.2044016
50. Guo G, Zhang J, Yorke-Smith N. A Novel Bayesian Similarity Measure for Recommender Systems. In: Twenty-third international joint conference on artificial intelligence, (2013).
51. Guo G, Zhang J, Thalmann D, Yorke-Smith N. Etaf: An Extended Trust Antecedents Framework for Trust Prediction. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). IEEE (2014). p. 540–7. doi:10.1109/asonam.2014.6921639
52. Chen C, Zhang M, Liu Y, Ma S. Social Attentional Memory Network: Modeling Aspect-And Friend-Level Differences in Recommendation. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (2019). p. 177–85.
53. Liang D, Krishnan RG, Hoffman MD, Jebara T. Variational Autoencoders for Collaborative Filtering. In: Proceedings of the 2018 world wide web conference (2018). p. 689–98. doi:10.1145/3178876.3186150
54. Ni J, Huang Z, Cheng J, Gao S. An Effective Recommendation Model Based on Deep Representation Learning. *Inf Sci* (2021) 542:324–42. doi:10.1016/j.ins.2020.07.038
55. Kim D, Park C, Oh J, Lee S, Yu H. Convolutional Matrix Factorization for Document Context-Aware Recommendation. In: Proceedings of the 10th ACM conference on recommender systems (2016). p. 233–40. doi:10.1145/2959100.2959165
56. Huang Z, Xu X, Ni J, Zhu H, Wang C. Multimodal Representation Learning for Recommendation in Internet of Things. *IEEE Internet Things J* (2019) 6(6):10675–85. doi:10.1109/jiot.2019.2940709
57. Liu J, Wu C, Wang J. Gated Recurrent Units Based Neural Network for Time Heterogeneous Feedback Recommendation. *Inf Sci* (2018) 423:50–65. doi:10.1016/j.ins.2017.09.048
58. Xiao J, Ye H, He X, Zhang H, Wu F, Chua T-S. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, (IJCAI-17). Melbourne, VIC, Australia (2017). p. 3119–25. doi:10.24963/ijcai.2017/435
59. Zeng B, Shang Q, Han X, Zeng F, Zhang M. Racmf: Robust Attention Convolutional Matrix Factorization for Rating Prediction. *Pattern Anal Applic* (2019) 22(4):1655–66. doi:10.1007/s10044-019-00814-2
60. Khattar D, Kumar V, Varma V, Gupta M. Hram: A Hybrid Recurrent Attention Machine for News Recommendation. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management (2018). p. 1619–22.
61. Zhang L, Liu P, Gulla JA. Dynamic Attention-Integrated Neural Network for Session-Based News Recommendation. *Mach Learn* (2019) 108(10):1851–75. doi:10.1007/s10994-018-05777-9
62. Tahmasebi H, Ravanmehr R, Mohamadrezaei R. Social Movie Recommender System Based on Deep Autoencoder Network Using Twitter Data. *Neural Comput Applic* (2021) 33(5):1607–23. doi:10.1007/s00521-020-05085-1
63. Behera DK, Das M, Swetanisha S. Predicting Users' Preferences for Movie Recommender System Using Restricted Boltzmann Machine. In: *Computational Intelligence in Data Mining*. Springer (2019). p. 759–69. doi:10.1007/978-981-10-8055-5_67
64. Polatidis N, GeorgiadisGeorgiadis CK, Pimenidis E, Mouratidis H. Privacy-preserving Collaborative Recommendations Based on Random Perturbations. *Expert Syst Appl* (2017) 71:18–25. doi:10.1016/j.eswa.2016.11.018
65. Salih Karakaşlı M, Aydin MA, Yarkan S, Ali B. Dynamic Feature Selection for Spam Detection in Twitter. In: International Telecommunications Conference. Springer (2019). p. 239–50.
66. Jain G, Sharma M, Agarwal B. Spam Detection in Social media Using Convolutional and Long Short Term Memory Neural Network. *Ann Math Artif Intell* (2019) 85(1):21–44. doi:10.1007/s10472-018-9612-z
67. Tajalizadeh H, Boostani R. A Novel Stream Clustering Framework for Spam Detection in Twitter. *IEEE Trans Comput Soc Syst* (2019) 6(3):525–34. doi:10.1109/tcss.2019.2910818
68. Zhao C, Xin Y, Li X, Zhu H, Yang Y, Chen Y. An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks. *Appl Sci* (2020) 10(22):8160. doi:10.3390/app10228160
69. Yang C, Harkreader R, Zhang J, Shin S, Gu G. Analyzing Spammers' Social Networks for Fun and Profit: a Case Study of Cyber Criminal Ecosystem on Twitter. In: Proceedings of the 21st international conference on World Wide Web (2012). p. 71–80.
70. Zhang Z, Hou R, Yang J. Detection of Social Network Spam Based on Improved Extreme Learning Machine. *IEEE Access* (2020) 8:112003–14. doi:10.1109/access.2020.3002940
71. Gao Y, Gong M, Xie Y, Qin AK. An Attention-Based Unsupervised Adversarial Model for Movie Review Spam Detection. *IEEE Trans multimedia* (2020) 23:784–96.
72. Quinn M, Olszewska JI. British Sign Language Recognition in the Wild Based on Multi-Class Svm. In: 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE (2019). p. 81–6.
73. An J, Cho S. Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability. *Spec Lecture IE* (2015) 2(1):1–18.
74. Zhao C, Xin Y, Li X, Yang Y, Chen Y. A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data. *Appl Sci* (2020) 10(3):936. doi:10.3390/app10030936
75. Chen C, Zhang J, Chen X, Yang X, Zhou W. 6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection. In: 2015 IEEE international conference on communications (ICC). IEEE (2015). p. 7065–70. doi:10.1109/icc.2015.7249453
76. Sze-To A, Wong AKC. A Weight-Selection Strategy on Training Deep Neural Networks for Imbalanced Classification. In: International Conference Image

Analysis and Recognition. Springer (2017). p. 3–10. doi:10.1007/978-3-319-59876-5_1
77. Alom Z, Carminati B, Ferrari E. *A Deep Learning Model for Twitter Spam Detection*, 18. Online Social Networks and Media (2020). p. 100079. doi:10.1016/j.osnem.2020.100079A Deep Learning Model for Twitter Spam Detection*Online Soc Networks Media*
78. Swe MM, Myo NN. Fake Accounts Detection on Twitter Using Blacklist. In: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). IEEE (2018). p. 562–6. doi:10.1109/icis.2018.8466499
79. Neha MV, Nair MS. A Novel Twitter Spam Detection Technique by Integrating Inception Network with Attention Based Lstm. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE (2021). p. 1009–14. doi:10.1109/icoei51242.2021.9452825
80. Arora S, Li Y, Liang Y, Ma T, Risteski A. A Latent Variable Model Approach to Pmi-Based Word Embeddings. *Tacl* (2016) 4:385–99. doi:10.1162/tacl_a_00106
81. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing Properties of Neural Networks. 2nd International Conference on Learning Representations, ICLR 2014-Conference Track Proceedings (2013), Banff, AB, Canada: Elsevier Inc.
82. Goodfellow IJ, Shlens J, Szegedy C. *Explaining and Harnessing Adversarial Examples* (2014). 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, United states: Elsevier Inc.
83. Luo Y, Boix X, Roig G, Poggio T, Qi Z. *Foveation-based Mechanisms Alleviate Adversarial Examples* (2015). arXiv preprint arXiv:1511.06292.
84. Gilmer J, Metz L, Faghri F, Schoenholz SS, Raghu M, Martin W, et al. *Adversarial Spheres* (2018). arXiv preprint arXiv:1801.02774.
85. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. *Adversarial Examples Are Not Bugs, They Are Features* (2019). arXiv preprint arXiv:1905.02175.
86. Yuan X, He P, Zhu Q, Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans Neural Netw Learn Syst.* (2019) 30(9):2805–24. doi:10.1109/tnnls.2018.2886017
87. Ji J, Du TY, Li JF, Shen C, Li B. Application of Artificial Intelligence Technology in English Online Learning Platform. *J Softw* (2021) 32(1): 41–9. doi:10.1007/978-3-030-89508-2_6
88. Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks. In: 2017 ieee symposium on security and privacy (sp). IEEE (2017). p. 39–57. doi:10.1109/sp.2017.49
89. Carlini N, Wagner D. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security (2017). p. 3–14.
90. Carlini N, Wagner D. *Magnet and" Efficient Defenses against Adversarial Attacks" Are Not Robust to Adversarial Examples* (2017). arXiv preprint arXiv:1711.08478.
91. Chen P-Y, Sharma Y, Zhang H, Yi J, Hsieh C-J. Ead: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. In: Thirty-second AAAI conference on artificial intelligence (2018).
92. Kurakin A, Goodfellow I, Bengio S. *Adversarial Examples in the Physical World* (2016).
93. Kurakin A, Goodfellow I, Bengio S. *Adversarial Machine Learning at Scale* (2016). arXiv preprint arXiv:1611.01236.
94. Madry A, Makelov A, Schmidt L, Tsipras D, Adrian V. *Towards Deep Learning Models Resistant to Adversarial Attacks*. In: 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings (2017), Vancouver, BC, Canada.
95. Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, et al. Boosting Adversarial Attacks with Momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018). p. 9185–93. doi:10.1109/cvpr.2018.00957
96. Moosavi-Dezfooli S-M, Fawzi A, Pascal F. Deepfool: a Simple and Accurate Method to Fool Deep Neural Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016). p. 2574–82. doi:10.1109/cvpr.2016.282
97. Baluja S, Fischer I. Learning to Attack: Adversarial Transformation Networks. In: Thirty-second aaai conference on artificial intelligence (2018).
98. Xiao C, Li B, Zhu J-Y, He W, Liu M, Song D. Generating Adversarial Examples with Adversarial Networks. In 2021 IEEE International Conference on Image Processing (ICIP). Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization (2018).
99. Bai T, Zhao J, Zhu J, Han S, Chen J, Li B, et al. Ai-gan: Attack-Inspired Generation of Adversarial Examples. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE (2021), 35(10):2543–7. doi:10.1109/icip42928.2021.9506278
100. Mao X, Chen Y, Wang S, Su H, Yuan H, Xue H. *Composite Adversarial Attacks* (2020). arXiv preprint arXiv:2012.05434.
101. Chen P-Y, Zhang H, Sharma Y, Yi J, Hsieh C-J. Zoo: Zeroth Order Optimization Based Black-Box Attacks to Deep Neural Networks without Training Substitute Models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security (2017). p. 15–26.
102. Ilyas A, Engstrom L, Athalye A, Lin J. Black-box Adversarial Attacks with Limited Queries and Information. In: International Conference on Machine Learning. Stockholm, Sweden: PMLR (2018). p. 2137–46.
103. Tim Salimans TH, Jonathan S, Chen X, Sidor S, Sutskever I. *Evolution Strategies as a Scalable Alternative to Reinforcement Learning* (2017). arXiv preprint arXiv:1703.03864.
104. Tu C-C, Ting P, Chen P-Y, Liu S, Zhang H, Yi J, et al. Autozoom: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks. *Aaai* (2019) 33:742–9. doi:10.1609/aaai.v33i01.3301742
105. Du J, Zhang H, Zhou JT, Yang Y, Feng J. *Query-efficient Meta Attack to Deep Neural Networks* (2019). arXiv preprint arXiv:1906.02398.
106. Yang B, Zeng Y, Jiang Y, Wang Y, Xia S-T, Guo W. Improving Query Efficiency of Black-Box Adversarial Attack. In: Computer Vision–ECCV 2020: 16th European Conference; August 23–28, 2020; Glasgow, UK. Springer (2020). p. 101–16.
107. Chen J, Jordan MI, Hopskipjumpattack MJW. A Query-Efficient Decision-Based Attack. In: 2020 ieee symposium on security and privacy (sp). IEEE (2020). p. 1277–94.
108. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical Black-Box Attacks against Machine Learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security (2017). p. 506–19. doi:10.1145/3052973.3053009
109. Zhou M, Wu J, Liu Y, Liu S, Zhu C. Dast: Data-free Substitute Training for Adversarial Attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020). p. 234–43. doi:10.1109/cvpr42600.2020.00031
110. Ma C, Cheng S, Chen L, Zhu J, Junhai Y. *Switching Transferable Gradient Directions for Query-Efficient Black-Box Adversarial Attacks* (2020). arXiv preprint arXiv:2009.07191.
111. Zhu Y, Cheng Y, Zhou H, Lu Y. Hermes Attack: Steal {DNN} Models with Lossless Inference Accuracy. In: 30th {USENIX} Security Symposium ({USENIX} Security 21) (2021).
112. Wang W, Yin B, Yao T, Zhang L, Fu Y, Ding S, et al. Delving into Data: Effectively Substitute Training for Black-Box Attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021). p. 4761–70. doi:10.1109/cvpr46437.2021.00473
113. Ma C, Chen L, Jun-Hai Y. Simulating Unknown Target Models for Query-Efficient Black-Box Attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021). p. 11835–44. doi:10.1109/cvpr46437.2021.01166
114. Das N, Shanbhogue M, Chen S-T, Hohman F, Li S, Chen L, et al. Shield: Fast, Practical Defense and Vaccination for Deep Learning Using Jpeg Compression. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018). p. 196–204.
115. Cheng Y, Wei X, Fu H, Lin S-W, Lin W. Defense for Adversarial Videos by Self-Adaptive Jpeg Compression and Optical Texture. In: Proceedings of the 2nd ACM International Conference on Multimedia in Asia (2021). p. 1–7. doi:10.1145/3444685.3446308
116. Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting Classifiers against Adversarial Attacks Using Generative Models. 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings, Vancouver, BC, Canada (2018).
117. Hwang U, Park J, Jang H, Yoon S, Cho NI. Puvae: A Variational Autoencoder to Purify Adversarial Examples. *IEEE Access* (2019) 7:126582–93. doi:10.1109/ACCESS.2019.2939352

118. Lin W-A, Balaji Y, Samangouei P, Chellappa R. *Invert and Defend: Model-Based Approximate Inversion of Generative Adversarial Networks for Secure Inference* (2019). arXiv preprint arXiv:1911.10291.
119. Zhang S, Gao H, Rao Q. Defense against Adversarial Attacks by Reconstructing Images. *IEEE Trans Image Process* (2021) 30:6117–29. doi:10.1109/tip.2021.3092582
120. Liu J, Zhang W, Zhang Y, Hou D, Liu Y, Zha H, et al. Detection Based Defense against Adversarial Examples from the Steganalysis point of View. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019). p. 4825–34. doi:10.1109/cvpr.2019.00496
121. Wang J, Zhao J, Yin Q, Luo X, Zheng Y, Shi YQ, et al. SmsNet: A New Deep Convolutional Neural Network Model for Adversarial Example Detection. *IEEE Trans Multimedia* (2021), IEEE.
122. Tian J, Zhou J, Li Y, Jia D. Detecting Adversarial Examples from Sensitivity Inconsistency of Spatial-Transform Domain. *Proc. AAAI Conf. Art. Intel.* (2021) 3511:9877–85. doi:10.1016/j.ins.2021.01.035
123. Chen J, Zheng H, Shangguan W, Liu L, Ji S. Act-detector: Adaptive Channel Transformation-Based Light-Weighted Detector for Adversarial Attacks. *Inf Sci* (2021) 564:163–92. doi:10.1016/j.ins.2021.01.035
124. Evan Sutanto R, Lee S. Real-time Adversarial Attack Detection with Deep Image Prior Initialized as a High-Level Representation Based Blurring Network. *Electronics* (2021) 10(1):52. doi:10.3390/electronics10010052
125. Liang B, Li H, Su M, Li X, Shi W, Wang X. Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction. *IEEE Trans Dependable Secure Comput* (2018) 18(1):72–85. doi:10.1109/TDSC.2018.2874243
126. Bai T, Luo J, Zhao J, Wen B, Wang Q. *Recent Advances in Adversarial Training for Adversarial Robustness* (2021). arXiv preprint arXiv:2102.01356.
127. Ali S, Najibi M, Amin G, Xu Z, John D, Studer C, et al. *Adversarial Training for Free!* (2019) Vancouver, BC, Canada, 32.
128. Zhang D, Zhang T, Lu Y, Zhu Z, Dong B. *You Only Propagate once: Accelerating Adversarial Training via Maximal Principle* (2019) Vancouver, BC, Canada, 32.
129. Wong E, Rice L, Kolter JZ. *Fast Is Better than Free: Revisiting Adversarial Training* (2020). arXiv preprint arXiv:2001.03994.
130. Gowal S, Dvijotham K, Stanforth R, Bunel R, Qin C, Uesato J, et al. *On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models* (2018). arXiv preprint arXiv:1810.12715.
131. Zhang H, Chen H, Xiao C, Gowal S, Stanforth R, Li B, et al. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. *Adv. Neural Inform. Process. Syst.* (2019) 2018(10495258):4939–4948.
132. Zhang H, Weng T-W, Chen P-Y, Hsieh C-J, Daniel L. Efficient Neural Network Robustness Certification with General Activation Functions. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual, United states: Association for Computational Linguistics (2018), 6066–80.
133. Gao J, Lanchantin J, Lou Soffa M, Qi Y. Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. In: 2018 IEEE Security and Privacy Workshops (SPW). IEEE (2018). p. 50–6. doi:10.1109/spw.2018.00016
134. Vijayaraghavan P, Roy D. Generating Black-Box Adversarial Examples for Text Classifiers Using a Deep Reinforced Model. *Mach. Learn. Knowl. Disc. Datab.* (2019). p. 711–726.
135. Iyyer M, John W, Gimpel K, Zettlemoyer L. Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. In: NAACL HLT 2018-2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference. New Orleans, LA, United states (2018). p. 1875–85.
136. Ren Y, Lin J, Tang S, Zhou J, Yang S, Qi Y, et al. Generating Natural Language Adversarial Examples on a Large Scale With Generative Models. *Front. Artif. Intell. App.* (2020) 325(09226389):2156–63. doi:10.3233/FAIA200340
137. Li J, Ji S, Du T, Li B, Wang T. Textbugger: Generating Adversarial Text against Real-World Applications. *IEEE Access* (2020) 8:79561–72.
138. Li L, Ma R, Guo Q, Xue X, Qiu X. *Bert-attack: Adversarial Attack against Bert Using Bert* (2020). arXiv preprint arXiv:2004.09984.
139. Jin D, Jin Z, Zhou JT, Szolovits P. Is Bert Really Robust? a strong Baseline for Natural Language Attack on Text Classification and Entailment. *Aaai* (2020) 34:8018–25. doi:10.1609/aaai.v34i05.6311
140. Alzantot M, Sharma Y, Ahmed E, Ho B-J, Srivastava M, Chang K-W. Generating Natural Language Adversarial Examples in. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. Virtual, United states (2018), 2890–96.
141. Cheng N, Chang G-Q, Gao H, Pei G, Zhang Y. Wordchange: Adversarial Examples Generation Approach for Chinese Text Classification. *IEEE Access* (2020) 8:79561–72.
142. Li D, Zhang Y, Peng H, Chen L, Brockett C, Sun M-T, et al. Contextualized Perturbation for Textual Adversarial Attack in. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2020), 5053–69.
143. Garg S, Bae GR. Bert-based Adversarial Examples for Text Classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020). p. 6174–81.
144. Maheshwary R, Maheshwary S, Pudi V. Generating Natural Language Attacks in a Hard Label Black Box Setting. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence (2021).
145. Yuan Z, Qi F, Yang C, Liu Z, Zhang M, Liu Q, et al. *Word-level Textual Adversarial Attacking as Combinatorial Optimization* (2019). arXiv preprint arXiv:1910.12196.
146. Yang G, Gong NZ, Cai Y. *Fake Co-visitation Injection Attacks to Recommender Systems*. NDSS (2017).
147. Fang M, Yang G, Gong NZ, Jia L. Poisoning Attacks to Graph-Based Recommender Systems. In: Proceedings of the 34th Annual Computer Security Applications Conference (2018). p. 381–92. doi:10.1145/3274694.3274706
148. Christakopoulou K, Banerjee A. Adversarial Attacks on an Oblivious Recommender. In: Proceedings of the 13th ACM Conference on Recommender Systems (2019). p. 322–30. doi:10.1145/3298689.3347031
149. Sun Y, Wang S, Tang X, Hsieh T-Y, Honavar V. Adversarial Attacks on Graph Neural Networks via Node Injections: A Hierarchical Reinforcement Learning Approach. *Proc Web Conf* (2020) 2020:673–83. doi:10.1145/3366423.3380149
150. Song J, Zhao L, Hu Z, Wu Y, Li Z, Li J, et al. Poisonrec: an Adaptive Data Poisoning Framework for Attacking Black-Box Recommender Systems. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE (2020). p. 157–68. doi:10.1109/icde48307.2020.00021
151. Chang H, Yu R, Xu T, Huang W, Zhang H, Cui P, et al. *Adversarial Attack Framework on Graph Embedding Models with Limited Knowledge* (2021). arXiv preprint arXiv:2105.12419.
152. Fang M, Gong NZ, Jia L. Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems. In: Proceedings of The Web Conference (20202020). p. 3019–25. doi:10.1145/3366423.3380072
153. Li B, Wang Y, Singh A, Vorobeychik Y. Data Poisoning Attacks on Factorization-Based Collaborative Filtering. *Adv Neural Inf Process Syst* (2016) 29:1885–93.
154. Lin C, Chen S, Li H, Xiao Y, Li L, Yang Q. Attacking Recommender Systems with Augmented User Profiles. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management (2020). p. 855–64. doi:10.1145/3340531.3411884
155. Fan W, Tyler D, Zhao X, Yao M, Liu H, Wang J, et al. Attacking Black-Box Recommendations via Copying Cross-Domain User Profiles. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE (2021). p. 1583–94. doi:10.1109/icde51399.2021.00140
156. Zhan H, Pei X. *Black-box Gradient Attack on Graph Neural Networks: Deeper Insights in Graph-Based Attack and Defense* (2021). arXiv preprint arXiv:2104.15061.
157. Ben F, Baskin C, Zheltonozhskii E, Alon U. *Single-node Attack for Fooling Graph Neural Networks* (2020). arXiv preprint arXiv:2011.03574.
158. Huang H, Mu J, Gong NZ, Qi L, Liu B, Xu M. *Data Poisoning Attacks to Deep Learning Based Recommender Systems* (2021). arXiv preprint arXiv:2101.02644.
159. Wu C, Lian D, Ge Y, Zhu Z, Chen E. Triple Adversarial Learning for Influence Based Poisoning Attack in Recommender Systems. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (2021). p. 1830–40. doi:10.1145/3447548.3467335
160. Pruthi D, Dhingra B, Lipton ZC. Combating Adversarial Misspellings with Robust Word Recognition in. ACL 2019-57th Annual Meeting of the

Association for Computational Linguistics, Proceedings of the Conference. Florence, Italy (2020), 5582–91.
161. Jia R, Raghunathan A, Göksel K, Liang P. Certified Robustness to Adversarial Word Substitutions in. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Hong Kong, China (2019), 4129–42.
162. Zhou Y, Jiang J-Y, Chang K-W, Wang W. Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification in. EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Hong Kong, China (2019), 4904–13.
163. Si C, Zhang Z, Qi F, Liu Z, Wang Y, Liu Q, et al. *Better Robustness by More Coverage: Adversarial Training with Mixup Augmentation for Robust fine-tuning* (2020). arXiv preprint arXiv:2012.15699.
164. Ren S, Deng Y, He K, Che W. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy (2019). p. 1085–97. doi:10.18653/v1/p19-1103
165. Wang Z, Liu H, Tang J, Yang S, Huang GY, Liu Z. Learning Multi-Level Dependencies for Robust Word Recognition. *Proc. AAAI Conf. Artif. Intell.* (2020) 34:9250–7. doi:10.1609/aaai.v34i05.6463
166. Shi Z, Zhang H, Chang K-W, Huang M, Hsieh C-J. Robustness Verification for Transformers in. *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics (2020), 164–171.
167. Ye M, Gong C, Liu Q. Safer: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Virtual, Online, United states (2020). p. 3465–75.
168. Mozes M, Stenetorp P, Bennett K, LewisGriffin D. Frequency-guided Word Substitutions for Detecting Textual Adversarial Examples in. EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. Virtual (2020), 171–86.
169. Zeng J, Zheng X, Xu J, Li L, Yuan L, Huang X. *Certified Robustness to Text Adversarial Attacks by Randomized [mask]* (2021). arXiv preprint arXiv: 2105.03743.
170. Wang W, Wang R, Ke J, Wang L. Textfirewall: Omni-Defending against Adversarial Texts in Sentiment Classification. *IEEE Access* (2021) 9: 27467–75. doi:10.1109/access.2021.3058278
171. Karimi A, Rossi L, Prati A. Adversarial Training for Aspect-Based Sentiment Analysis with Bert. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021). p. 8797–803. doi:10.1109/icpr48806.2021.9412167
172. Hu X, Liu B, Shu L, Philip SY. Bert post-training for Review reading Comprehension and Aspect-Based Sentiment Analysis in. NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. Minneapolis, MN, United states (2019), 2324–35.
173. Miyato T, Dai AM, Goodfellow I. Adversarial Training Methods for Semi-supervised Text Classification in. 5th International Conference on Learning Representations, ICLR 2017-Conference Track Proceedings. Toulon, France (2016).
174. Du Y, Fang M, Yi J, Xu C, Cheng J, Tao D. Enhancing the Robustness of Neural Collaborative Filtering Systems under Malicious Attacks. *IEEE Trans Multimedia* (2018) 21(3):555–65. doi:10.1109/tmm.2018.2887018
175. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. In: 2016 IEEE symposium on security and privacy (SP). IEEE (2016). p. 582–97. doi:10.1109/sp.2016.41
176. Tang J, Du X, He X, Yuan F, Qi T, Chua T-S. Adversarial Training towards Robust Multimedia Recommender System. *IEEE Trans Knowledge Data Eng* (2019) 32(5):855–67. doi:10.1109/TKDE.2019.2893638
177. Manotumruksa J, Yilmaz E. Sequential-based Adversarial Optimisation for Personalised Top-N Item Recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020). p. 2045–8. doi:10.1145/3397271.3401264
178. Li R, Wu X, Wang W. Adversarial Learning to Compare: Self-Attentive Prospective Customer Recommendation in Location Based Social Networks. In: Proceedings of the 13th International Conference on Web Search and Data Mining (2020). p. 349–57.
179. Wang J, Han P. Adversarial Training-Based Mean Bayesian Personalized Ranking for Recommender System. *IEEE Access* (2019) 8:7958–68.
180. Shahrasbi B, Mani V, Arrabothu AR, Sharma D, Kannan A, Kumar S. *On Detecting Data Pollution Attacks on Recommender Systems Using Sequential gans* (2020). arXiv preprint arXiv:2012.02509.
181. Wu C, Lian D, Ge Y, Zhu Z, Chen E, Yuan S. Fight Fire with Fire: Towards Robust Recommender Systems via Adversarial Poisoning Training. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021). p. 1074–83. doi:10.1145/3404835.3462914
182. Yi Q, Yang N, Yu P. Dual Adversarial Variational Embedding for Robust Recommendation. *IEEE Trans Knowledge Data Eng* (2021). doi:10.1109/tkde.2021.3093773
183. Shi Z, Huang M. *Robustness to Modification with Shared Words in Paraphrase Identification* (2019). arXiv preprint arXiv:1909.02560.
184. PangKoh W, Liang P. Understanding Black-Box Predictions via Influence Functions. In: International Conference on Machine Learning. Sydney, NSW, Australia: PMLR (2017). p. 1885–94.
185. Xue M, Yuan C, Wang J, Liu W. Dpaeg: A Dependency Parse-Based Adversarial Examples Generation Method for Intelligent Q&a Robots. *Security and Communication Networks*Hindawi (2020). p. 2020.
186. Deng E, Qin Z, Meng L, Ding Y, Qin Z. Attacking the Dialogue System at Smart home. In: International Conference on Collaborative Computing: Networking, Applications and Worksharing. Springer (2020). p. 148–58.
187. Ebrahimi J, Daniel L, Dou D. Efficient Neural Network Robustness Certification with General Activation Functions in. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics (2018), 653–63.
188. Wang J, Xu C, Guzmán F, El-Kishky A, Tang Y, Rubinstein BIP, et al. Putting Words into the System's Mouth: A Targeted Attack on Neural Machine Translation Using Monolingual Data Poisoning *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (2021), 1463–73.
189. Cheng Y, Jiang L, Macherey W. Robust Neural Machine Translation with Doubly Adversarial Inputs in. ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. Florence, Italy (2019), 4324–33.
190. Cheng Y, Jiang L, Macherey W, Jacob E. Advaug: Robust Adversarial Augmentation for Neural Machine Translation in. Proceedings of the Annual Meeting of the Association for Computational Linguistics. Virtual, United states (2020), 5961–70. doi:10.18653/v1/2020.acl-main.529

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.