https://doi.org/10.33919/JCAL.23.1.1

TEXT DATA AUGMENTATION USING GENERATIVE ADVERSARIAL NETWORKS – A SYSTEMATIC REVIEW

Kanishka Silva[1*], Burcu Can[2], Raheem Sarwar[3], Frederic Blain[4], Ruslan Mitkov[1]

[1] University of Wolverhampton, Wolverhampton, United Kingdom

[2] Department of Computing Science and Mathematics, University of Stirling, Stirling, United Kingdom

[3] Department of Operations, Technology, Events and Hospitality Management, Faculty of Business and Law, Manchester Metropolitan University, United Kingdom

[4] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands

[*] Corresponding author. Email: kanishka.silva@wlv.ac.uk

Abstract

Insufficient data is one of the main drawbacks in natural language processing tasks, and the most prevalent solution is to collect a decent amount of data that will be enough for the optimisation of the model. However, recent research directions are strategically moving towards increasing training examples due to the nature of the data-hungry neural models. Data augmentation is an emerging area that aims to ensure the diversity of data without attempting to collect new data exclusively to boost a model's performance.

Limitations in data augmentation, especially for textual data, are mainly due to the nature of language data, which is precisely discrete. Generative Adversarial Networks (GANs) were initially introduced for computer vision applications, aiming to generate highly realistic images by learning the image representations. Recent research has focused on using GANs for text generation and augmentation. This systematic review aims to present the theoretical background of GANs and their use for text augmentation alongside a systematic review of recent textual data augmentation applications such as sentiment analysis, low resource language generation, hate speech detection and fraud review analysis. Further, a notion of challenges in current research and future directions of GAN-based text augmentation are discussed in this paper to pave the way for researchers especially working on low-text resources.

Keywords: *Text Data Augmentation, Generative Adversarial Networks, Adversarial Training, Text Generation*

1. Introduction

Computational models in deep learning and machine learning usually perform better when high-quality and balanced datasets are available in natural language processing applications. However, it is usually challenging to obtain a high-quality dataset; for instance, in supervised learning tasks, we often need to deal with the lack of labelled data or a limited amount of labelled data, which directly affects the model's performance. Obtaining a large-scale dataset is time-consuming and associated with a higher cost. Therefore, expanding a given smaller dataset artificially for any natural language processing task is a promising solution. Applying data augmentation for NLP tasks, specifically for text-based applications, may exhibit lower accuracies due to language-variant characteristics such as grammatical structure. For instance, according to Luo et al. (2021), a text classification task would fail to improve performance due to grammatical errors or uncontrolled sentiment characteristics in the generated text. Although we need more data in data augmentation, replicating data is not a solution, as it will eventually lead to model overfitting.

Generative Adversarial Networks (Goodfellow et al., 2014) aim to synthesise real-world data as closely as possible. As improvements to the original GAN model proposed by Goodfellow et al., several other studies stabilised GAN training along with different loss functions (Nowozin et al., 2016; Mao et al., 2017; Arjovsky and Bottou 2017). Several other notable GAN architectures are Conditional Generative Adversarial Networks (Mirza and Osindero, 2014), Deep Convolutional Generative Adversarial Networks (Radford et al., 2018), Coupled Generative Adversarial Networks (Liu and Tuzel, 2016), Cycle-Consistent Generative Adversarial Networks (Zhu et al., 2017) and Information Maximizing Generative Adversarial Networks (Chen et al., 2016). Given the objective of GAN models, generating new data while being closer to the original data distribution is feasible to apply for data augmentation.

This paper aims to pave the way for researchers especially working on low textual resources, by reviewing previous work in textual data augmentation using GAN models in various NLP application domains. In this sense, this paper is the first systematic review focusing on GAN-based text data augmentation. Furthermore, we surveyed text augmentation application domains such as sentiment analysis, hate speech detection, low resource language generation and fraud text identification.

The research questions for this systematic study are as follows:

1. How can text augmentation help to improve a computational model's performance?
2. How can GAN models be utilised for text data augmentation?
3. What are the challenges in GAN-based text augmentation worth addressing in future research?

The rest of the paper is structured as follows: Section 2 describes the methodology followed for the systematic review and paper screening, such as inclusion and exclusion criteria. Section 3 briefly introduces data augmentation, and Section 4 presents a comprehensive overview of Generative Adversarial Networks. Section 5 systematically reviews a few applications using GAN-based text augmentation. Section 6 summarises text data augmentation challenges and potential future directions. Finally, Section 7 summarises the objectives of this study.

2. Methods

This systematic review adheres to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009). We filtered the articles through a well-defined inclusion-exclusion strategy per the PRISMA guidelines following through the identification, screening, exclusion, and inclusion stages. Figure 2.1 shows the PRISMA flowchart we used with filtered paper counts in each stage.

We conducted the search initialisation as per the PRISMA guidelines (Moher et al., 2009) and collected articles from digital libraries such as Scopus, Web of Science, IEEE Xplore, Science Direct, Google Scholar and Semantic Scholar, which were published between 2017 and 2022, with a search duration spanning from March 2022 to May 2022. We used some keywords to search the databases. Initially, we used key phrases such as "text data augmentation using generative adversarial networks" and "text augmentation using GAN". We then narrowed the search to the scope of applications, such as "Generative Adversarial Network data augmentation for fraud text identification" and "low resource language generation using GANs". Further, we utilised complex search strings to combine similar keywords with AND and different keywords with OR. For instance, "text augmentation" AND "text synthesis" and "text augmentation for low resource languages" OR "synthesised text in semantic analysis". Altogether we collected 257 papers initially and removed 96 duplicate entries, resulting in 161 papers for the screening stage.

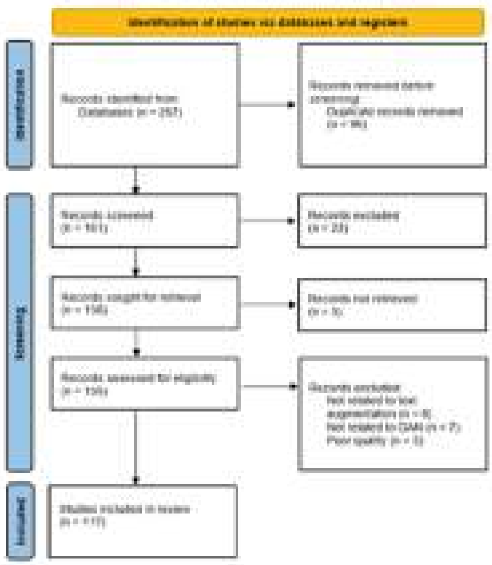

Figure 2.1: The PRISMA guideline flowchart used in this review (Moher et al., 2009)

Twenty-three articles were excluded during the screening process upon careful scan through the title and abstract. Then another exclusion step was performed considering full-text availability, which excluded three papers from the results. In the final step in screening, we considered whether the selected papers aligned with the stated research questions. We excluded 17 papers since they were unrelated to text augmentation or GAN, and some had poor-quality content. A total of 117 articles were selected eventually, and the distribution is illustrated in Figure 2.2. Finally, the papers were grouped hierarchically for a clear presentation in the review. Several papers were included during the write-up period since those papers were vital in explaining the theoretical background.

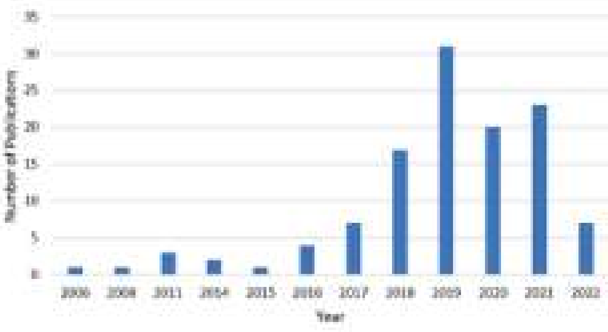

Figure 2.2: Numbers of selected publications over the years

3. Data Augmentation

Data augmentation generates a massive amount of data from a given small set of available data, guaranteeing an increased model accuracy. The simplicity of the proposed data augmentation approaches is a must to replace the time-intensive and cost-ineffective manual data collection and annotation to increase the size of an existing small-scale dataset. Feng et al. (2021) claim that a simple augmentation approach and accuracy boosting are trade-offs in data augmentation because overfitting will occur if the generated data is too identical to the original one. Therefore, the augmented data should be similar but deviate from the original data distribution. A typical approach is to perform data augmentation before the training is conducted and then mix the augmented data with the existing training data for training purposes. Another approach is generating data while the training occurs, a common technique in GAN-based data augmentation, especially in computer vision applications.

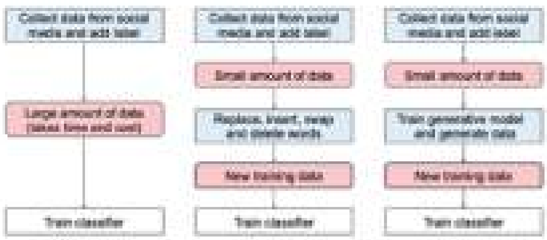

Figure 3.1: The methods used for collecting training data for a classifier. Left to right: a general method, dictionary-based data augmentation, generative model-based data augmentation (Luo et al., 2021)

Recent trends in NLP applications are heading towards leveraging large pre-trained models, especially in low-resource domains. Due to the exploration of new tasks, more data is the primary demand, but it is costly and time-intensive to annotate a large set of training data manually. Since high-quality data ensures the model's accuracy in conventional NLP approaches, it is difficult to turn a blind eye to this research gap. Moreover, low-resource scenarios, such as low-resource language data generation, also require a decent amount of training data. In such cases, augmenting data artificially is quite reasonable and adequate.

Overall, three techniques are used in data augmentation rule-based, example-interpolation-based and model-based (Feng et al., 2021). Rule-based approaches either consider the model's feature space (Xie et al., 2020; Wei and Zou 2019; Paschali et al., 2019) or use a graphical representation of the individual sentences (Chen et al., 2020; Şahin and Steedman, 2018). The example-interpolation technique takes two or more real examples and then alters the input and output labels. MIXUP architecture (Zhang et al., 2018) which follows the example-interpolation technique, has been later developed into different variations. Such variations are CUTMIX (Yun et al., 2019), which mixes two selected example images by replacing small sub-regions and Seq2MIXUP (Guo 2020), which generalises MIXUP for the sequence transduction task. Model-based techniques use sequence-to-sequence (seq2seq) models (Kumar et al., 2019; Sennrich et al., 2016) and language models based on recurrent neural networks and transformers (Sennrich et al., 2016; Yang et al., 2020).

Several data augmentation approaches in NLP include facilitating low-resource languages such as Turkish, Nepali, and Sinhala (Fadaee et al., 2017; Qin et al., 2021), bias mitigation (Zhao et al., 2018; Lu et al., 2020) and adversarial training (Jia et al., 2019; Kang et al., 2018). Moreover, applied NLP tasks that use data augmentation for performance gain involve classification (Wei and Zou 2019; Chen et al., 2020; Anaby-Tavor et al., 2020), summarisation (Fabbri et al., 2021; Parida and Motlicek 2019; Zhu et al., 2022), question answering (Longpre et al., 2019; Yang et al., 2019; Riabi et al., 2021), and dialogue systems (Quan and Xiong 2019; Louvan and Magnini 2020; Hou et al., 2018; Kim et al., 2019).

Initial approaches in textual data augmentation involve replacing words with synonyms or removing random words (Wei and Zou, 2019), which is not promising because of minor accuracy improvements due to overfitting, mainly in classification tasks. The data augmentation strategies followed for the textual data fall into three main categories: dictionary-based data augmentation, generative model-based, and general method, as in Figure 3.1 (Luo et al., 2021). Wei and Zou (2019) proposed a data augmentation strategy for text classification using a synonym dictionary to randomly increase the number of data points by inserting, replacing, deleting and swapping a word in a sentence. However, the performance with the synonym dictionary method (Wei and Zou, 2019) drops when the original data changes by more than a 10% ratio. Such approaches often exhibit the limitation of retaining sentiment information and even result in a drastic change in the actual sentiment class (Luo et al., 2021).

Generative models align with the probability distribution of the training data upon new data generation. Given that text generation is a complex task, such approaches were not entirely promising in text-based applications, specifically in classification models (Luo et al., 2021). Several generative models based on data augmentation were proposed by Anaby-Tavor et al. (2020), Feng et al. (2020), Radford et al. (2019). Apart from these text-generation strategies for text augmentation, generative adversarial networks are gaining popularity due to generating similar but fake data. Most data augmentation applications using GANs are in the computer vision area. However, there has been an increasing interest in using GANs for text data augmentation in the last few years.

4. Generative Adversarial Networks (GANs)

Machine learning models can be categorised into generative models and discriminative models. The discriminative models involve classification tasks that aim to predict the class labels by modelling a given feature set of inputs. In generative models, given the class and introduced noise, the distribution of the feature set is generated. Goodfellow et al. (2014) introduced a powerful generative model, Generative Adversarial Networks (GANs), adhering to a minimax game of two competing networks. The GAN model's main components compose a generator similar to a decoder and a discriminator that functions as a classifier. GANs have produced high-quality and diverse images for data augmentation in computer vision applications. Several GAN models which address image data are: face generation using StyleGAN (Karras et al., 2019), image translation using CycleGAN (Zhu et al., 2017), transforming doodles into pictures using GauGAN (Park et al., 2019) and generating 3D images using 3D-GAN (Wu et al., 2016), Wasserstein-GAN (Arjovsky and Bottou, 2017), coupled-GAN (Liu and Tuzel, 2016) and StackGAN (Zhang et al., 2017). The underpinning theories with these GAN applications deviate from the text data generation using GANs in minor aspects, but the intuition is the same by adhering to generator-discriminator architecture.

In GAN architecture, generator G learns to create fake samples that resemble real examples, and discriminator D learns to distinguish real samples from fake samples. The generator model is not sophisticated at the beginning to allow stable training. The discriminator mimics a classifier's behaviour. The probability outputs generated by the discriminator serve as an input for the generator. Both generator and discriminator are based on two separate neural networks. Figure 4.1 illustrates a GAN architecture. The input to the generator model is random noise, and the outputs are also randomly generated noisy samples. The generator expects to be as primitive as possible at this stage. Then the output is tuned with the response obtained from the discriminator. The generated samples become closer to the original data instances as the training continues. Following a minimax game theory, the generator and discriminator act as opponents trying to fool each other, eventually increasing the GAN model's performance on a particular task. The discriminator takes both original samples and the feature distribution of generated fake samples to classify both samples. Finally, when the discriminator cannot perform the classification correctly anymore, it is the point where the generator starts to make new samples which do not exist in the training data. Applications of GANs include super-resolution, assisting artists and element abstraction, specifically in the image domain.

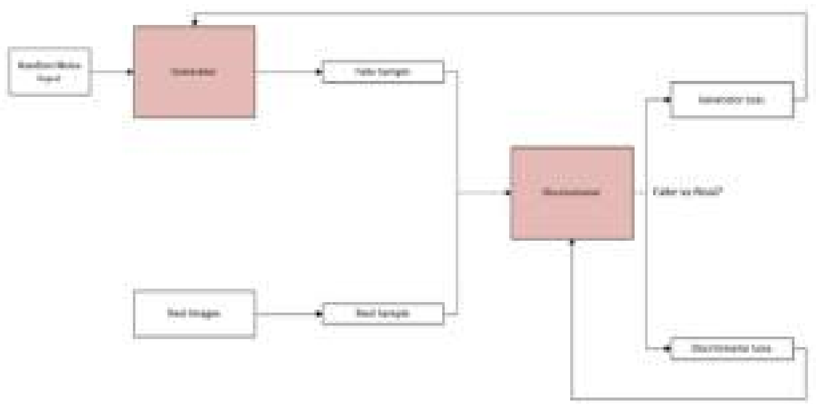

Figure 4.1: GAN Architecture

GAN models use adversarial concepts of producing fake samples mimicking real ones. The overall model improves continuously until an equilibrium point is reached due to competitive training of both the generator and discriminator. This concept is called the Minimax game, a decision rule with alternate moves for both players. Only one player wins by maximising their win in this concept, while the other tries to minimise the loss. Borrowing this idea for the GAN model, the generator tries to minimise the probability output of the discriminator, which is labelled as 'fake'. Simultaneously, the generator maximises the probability of classifying real and fake samples.

Equation (1) mathematically defines the minimax game of a GAN model: G is the generator, D is the discriminator, x denotes the real sample input, and D(x) is the probability of the label for the real sample. While z is the noise or the latent space vector used to provide inputs to the generator, G(z) indicates generated fake samples. The discriminator outputs that are expected for these two classes, respectively, are G(x) = 1 and D(G(z)) = 0. Mainly, the objective of the generator is to make the discriminator identify fake samples as real ones, i.e., D(G(z)) = 1, which results in minimising 1-D(G(z)):

$$\min_{G} \max_{D} V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[1 - \log D(G(z))] \quad (1)$$

When training the generator to minimise 1-D(G(z)), the generator's output should collectively provide input to the discriminator. Then the discriminator's loss should be backpropagated into the generator. To pass the loss gradients back to the generator, the selection criteria within the generator should be a differentiable function.

If we consider an RNN-based text generator, the next word in a sentence generated at each time step corresponds to the one with maximum probability in the softmax distribution. Suppose the GAN generator is implemented using a similar RNN to generate texts. However, the corresponding picking function is non-differentiable in the GAN generator. This issue does not apply to continuous data such as images. Using GANs for text generation is challenging due to the nature of textual data, which does not involve continuous and numerical data. However, since the text does not carry any of these features, despite the challenges, the following approaches were introduced to utilise GANs for text generation: the reinforcement algorithm-based method (Yu et al., 2017), the Gumbel-softmax approximation method (Kusner and Hernández-Lobato, 2016) and the method of avoiding discrete spaces (Donahue and Rumshisky, 2018).

Using reinforcement learning is presented by Fedus et al. (2018) and Yu et al. (2017). Suppose text generation is performed via a Reinforcement Learning (RL) agent, where the agent generates the next word based on the current state s, the previously generated sentence. A word vocabulary is used to define the action set. A reward is received once the RL agent reaches the end of the sentence action. In GAN architecture, the discriminator returns the overall reward.

Given the start state S_0, ϕ-parameterised discriminator model D_ϕ, sequence to produce $Y_{1:T} = (y_1, \ldots, y_t, \ldots, y_T)$, current state $s = Y_{1:t-1}$ and the reward for a complete sentence RT, the θ-parameterised generator model G_θ, a gradient method is utilised to find the optimal parameters θ^* by applying gradient descent as follows:

$$\theta \leftarrow \theta + \alpha_h \nabla_\theta J(\theta) \qquad (2)$$

while maximising the overall reward as given below:

$$J(\theta) = \Sigma_{y_1 \epsilon Y} G_\theta(y_1|s_0) Q_{D\phi}^{G\theta}(s_0, y_1) \qquad (3)$$

A discriminator network performs classification on input sentences by providing a metric of how real it is. G represents parametrised policy $\pi(a|s,\theta)$ which takes a set of words as input to produce a probability distribution for the next word. During the training process, Monte-Carlo rollouts calculate an intermediate reward, and the discriminator provides the reward for the entire sentence. Persisting issues with this method include high variance in gradient estimate with each episode, resulting in an unstable training process and slow convergence. Pretrained generator and discriminator models can speed up training to solve these problems. Another problem also occurs when the state-action space is vast; for example, with an extensive vocabulary set, it tends to converge to local minima.

Due to the issues mentioned earlier with the Reinforcement Learning approach, recent research focuses on investigating other solutions for discrete data generation using GAN models. Selecting the next word in text generation maximises the probability generated via the softmax function at each time step. This selection operation is non-differentiable. Suppose the output y is a one-hot-vector with |V|-dimensions and h hidden states. Then the sampling is performed as follows:

$$p = softmax(h) \qquad (4)$$

Another sampling method is to use a vector of samples g from a Gumbel distribution as follows:

$$y = one\_hot(arg\,max_i(h_i + g_i) \qquad (5)$$

To make the argmax() function differentiable, a softmax approximation and an additional temperature parameter τ are introduced as given below:

$$y = softmax(1/\tau(h+g)) \qquad (6)$$

so that when $\tau \to 0$, the output distribution converges to a one-hot vector. During the training, τ is initialised with larger values, which converge on zero, as mentioned in Kusner and Hernández-Lobato (2016) and Donahue and Rumshisky (2018).

In encoder-decoder mapping, the encoder projects the input space onto a smaller dimensionality, and the decoder reconstructs the input from this representation. The solution for GAN text generation is not to consider it a separate discrete token generation. Instead of decomposing a given input sequence of discrete word tokens, this approach works with continuous space vectors, which are not human-readable. The problem arises in the discriminator's input representation while feeding the real sentences, which the auto-encoder facilitates. At the end of the training, the generator network outputs sentence vectors.

5. GAN for Text Data Augmentation

GANs have already been used for text data augmentation for various NLP applications listed below. However, before reviewing such NLP applications, it is noteworthy to mention GAN models' drawbacks in classification tasks such as sentiment analysis. For example, GANs may generate augmented data in opposite polarity, drastically impacting a sentiment analysis task. Nevertheless, GAN-based data augmentation can mitigate class imbalance problems by generating missing class data with controlled generation. Moreover, in the tasks such as bot-generated data identification, GAN-based fake data generation provides a promising adversarial approach. Collecting and analysing such datasets manually in practical cases is difficult.

5.1 Applications

Many NLP applications have used GANs for text data augmentation. These NLP applications include sentiment analysis, hate speech detection, low resource language generation, fraud detection, and code-switching sentence generation.

5.1.1 Sentiment Analysis

The challenges in sentiment analysis include a lack of data for low-resource languages and an imbalance issue in available datasets. Transfer learning (Gupta et al., 2018) and semi-supervised learning (Goldberg and Zhu, 2006) are alternatives in low-resource scenarios, but text-generation models also facilitate such problems. As mentioned in (Gupta, 2019), several techniques were introduced for sentiment analysis in low-resource scenarios, such as semi-supervised learning (Socher et al., 2011), regularisation methods (Gupta et al., 2018; Sindhwani and Melville, 2008) and latent variable models (Täckström and McDonald, 2011).

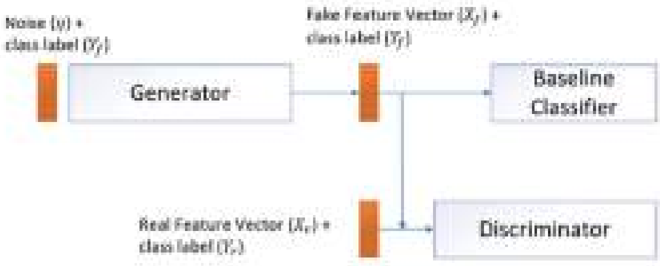

Figure 5.1: cGAN architecture (Gupta, 2019)

A variation of conditional GAN for low-resource datasets was introduced by Gupta (2019) with a baseline classifier in place apart from the generator and discriminator model. The implementation follows three approaches to ensure convergence: model pretraining from an available large dataset, input noise addition, and one-sided label smoothing, as illustrated in Figure 5.1. Both generator and discriminator employ feed-forward neural networks. The baseline classifier is pre-trained on a target task dataset and uses a shallow neural network architecture. The cross-entropy loss is used to learn the discriminator parameters as follows:

$$L_D = -y \log(D([x_r;y_r])) - (1-y)log(1-D([x_f;y_f])) \quad (7)$$

Here, each $[x_f; y_f]$ represents the concatenation with a label representation yf while assigned probabilities at discriminator are denoted by $D([x_r;y_r])$ and $D([x_f;y_f])$. Two generator losses are combined as given below:

$$L_G = L_{G1} + \lambda L_{G2} \quad (8)$$

where $L_{G1} = -\log(D([x_f; y_f]))$; $xf = G(\eta)$; $L_{G2} = -CE(y_f, C(x_f))$ (9)

The standard generator loss L_{G1} is to fool the discriminator while L_{G2} is to handle cross-entropy loss on the base classifier with λ hyper-parameter. $G_{(\eta)}$ corresponds to the generated output x_f with noise input η (Gupta, 2019).

Evaluation in Gupta (2019) is performed on the base classifier C_b, cGAN classifier C_f and a classifier on Twitter data C_t. Due to the discriminative power of generated data, C_f performs better, and the accuracy of C_t is mainly due to knowledge transfer. The evaluation of movie and product reviews has shown a significant accuracy increase of 1.76% and 1.7%, respectively, compared to the base classifier, which only uses actual data without utilising the generated data. As shown in Figure 5.2, T-SNE distribution and the projection of real vs fake data reveal that the generated data does not cover real data's entire feature space. Further, it is not easy to find a massive pre-trained dataset for the data augmentation task. Future directions include selective data generation in smaller spaces.

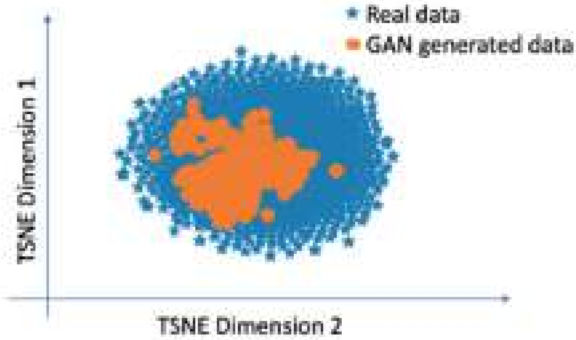

Figure 5.2: Real and fake data distribution, as observed on a 2-D projection of data points obtained using the t-SNE method (Gupta, 2019)

Another issue in sentiment analysis is the training on long texts in a low-resource dataset. As mentioned before, text generation models are prone to generating inaccurate sentiment information for the generated texts. Luo et al. (2021) propose a penalty-based SeqGAN for generating high-quality long-text data improving the SeqGAN model (Yu et al., 2017). The main challenge in using long text data is the low accuracy obtained when using such long text data in a classifier. The works of Luo et al. (2021) present an LSTM model with attention which performs sentence compression for the given training data. A sentiment dictionary aids in addressing the issue of losing sentiment words during the compression. With RL to address discrete data issues, the generator produces sentence sequence s based on the x token of the real word. The GAN model consists of a parameterised generator $G(\theta_g)$ and a discriminator $D(\theta_d)$ that aim to maximise the reward $G(x|s;\theta_g)D(x;\theta_d)$:

$$J_G(x) = \begin{cases} \mathbb{E}_{x\sim p_g}[-\log(D(x;\theta_d))] \\ \mathbb{E}_{x\sim p_g}[-\log(G(x|s;\theta_g)D(x;\theta_d))] \\ \mathbb{E}_{x\sim p_g}[G(x|s;\theta_g)V(x)] \end{cases} \quad (10)$$

The applied penalty-based objective on the generator is forced to minimise the overall penalty $G(x|s;\theta_g)V(x)$ given that $V(x) = 1 - D(x;\theta_d)$, which leads to generating grammatically correct sentences.

Compared to the previous cGAN model (Gupta, 2019), this model requires no pre-training step with another dataset on the target task. The evaluation parameters involve classification accuracy, usability, novelty, and the diversity of the generated data, which outperforms the state-of-the-art accuracy (Wei and Zou, 2019).

5.1.2 Hate Speech Detection

Hate speech detection is usually performed by supervised models. However, most of the available datasets are imbalanced, which is one reason for the low performance of the hate detection models. Applying data augmentation for the class with fewer examples is a reasonable solution, but this is a challenging task for text generation. Cao and Lee (2020) introduce HateGAN, a GAN model aiming for hate speech detection using a deep generative RL model based on hateful tweets. The overall architecture is illustrated in Figure 5.3. The model adopts SeqGAN (Yu et al., 2017) by adding a toxicity scorer (Figure 5.4), which is pre-trained as a multi-label classifier to provide realistic scores and hate scores.

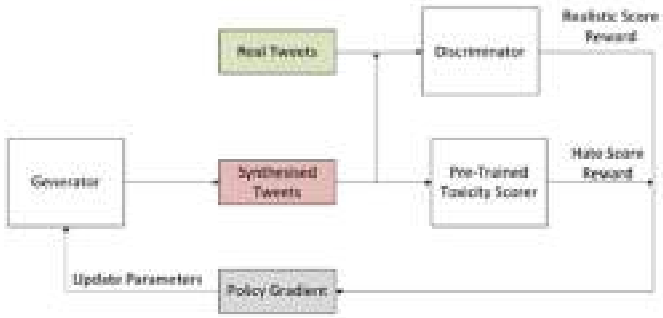

Figure 5.3 Architecture of the HateGAN model (Cao and Lee, 2020)

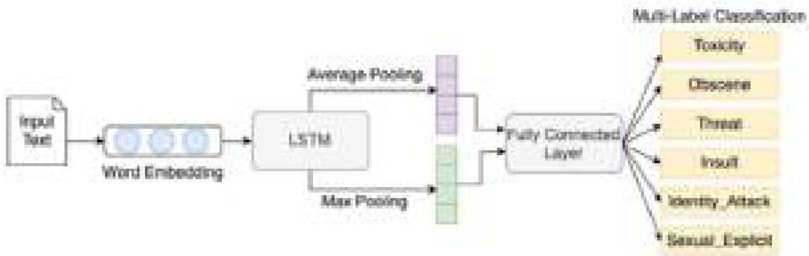

Figure 5.4: Toxicity scorer that is pre-trained as a multi-label classification model (Cao and Lee, 2020)

Given that S is a scoring module, N is the number of Monte Carlo searches, and x_i is the i-th Monte Carlo result, the expected reward from a sentence which is an action value for selecting the t-th word w_t is computed as follows:

$$r(state = (w_1, \ldots, w_{t-1}), actions = w_t) = \frac{1}{N}\sum_{i=1}^{N}(S(x_i)) \quad (11)$$

The loss as a negative expected reward is defined as follows:

$$\begin{aligned} Loss(\alpha) &= -\sum_{t=1}^{n} \mathbb{E}_{[w_{1:t-1}]\sim G_\alpha}\left[\mathbb{E}_{w_t\sim G_\alpha}[r(w_t)]\right] \\ &\approx -\sum_{t=1}^{n} \sum_{w_t\in V} G_\alpha(w_t|w_{1:t-1}) \frac{1}{N} \sum_{i=1}^{N} S(x_i^t) \end{aligned} \quad (12)$$

The final combined reward becomes:

$$r(x) = Discriminator(x) + \sigma ToxicityScorer(x) \quad (13)$$

where x is the input sentence and σ is a hyperparameter.

5.1.3 Low Resource Language Generation

Question Answering (QA) is useful in deep learning since many deep learning applications can be modelled as QA problems. Developing a QA system in a low-resource language is challenging due to insufficient annotated datasets. For instance, according to Sun et al. (2019), a low-resource language, Tibetan demonstrates challenges in building such a question-answering model because of the language features such as longer sentences, complex syntactic structures and strict grammatical rules. Sun et al. (2019) introduce QuGAN, using Quasi-Recurrent Neural Networks (QRNN) and Reinforcement Learning as a QA corpus generation model for the Tibetan language. QRNN consists of convolution components to extract features followed by an f-pooling component with a forget-gate to reduce the dimension of the features. The use of LSTM and CNN in the generator enables addressing the issue of processing longer sequences and parallel execution. The random initialisation of questions with Maximum Likelihood Estimation (MLE) ensures that both generated and original data follow a closer probability distribution.

Further optimisation proposes a reward strategy and Monte Carlo Search Strategy in the Reinforcement Learning model, which involves predicting the next sentence score based on the partially generated sequence rather than using the entire text. Following that, a BERT model facilitates the correction of the grammar of the generated text. The model evaluation uses data collected from the Tibetan website that involves 21783 questions for training different models with SeqGAN as the base model, QuGAN, QuGAN without Monte Carlo optimisation, QuGAN with BERT but without Monte Carlo Optimisation and QuGAN with BERT. QuGAN (Sun et al., 2019) has proven improvement of BLEU-2 score by 13.07 compared to the baseline with notable speed improvements. Further improvements can be made by generating grammatically correct questions by incorporating Tibetan grammar information and adding argument functions.

Another low-resource language scenario are the tasks involving regional dialects. A modified SentiGAN (Wang and Wan, 2018) based model (Carrasco et al., 2021) introduces an approach for data augmentation for Arabic Regional Dialects. Given that existing rich-annotated Dialectal Arabic datasets exhibit data scarcity, text data augmentation is also a solution for this issue. The selected regional Arabic dialects in that study are Egypt, Gulf, Maghreb, Levant, and Iraq. The generator uses an LSTM model with a policy gradient and a distractor using a CNN. Although the traditional SentiGAN (Wang and Wan, 2018) incorporates two sentiments, five dialects are generated using five generator/discriminator sets here. The model deviates from the other GAN-based text data augmentation models with a penalty instead of a reward for the discriminator model. The model generates a higher number of sentences than the original data size but with a reduced vocabulary size due to the usage of only the common words. The MADAR dataset is used for training and evaluating based on two new metrics to measure the novelty and diversity of the augmented texts and to assess further on four classification scenarios. Further improvement was also made by Wang and Wan (2018) by augmenting country-level dialects for Dialectical Arabic datasets.

In multilingual communities, loanwords are defined as words introduced and adopted from another language. Mi et al. (2021) provide data augmentation methodology to improve such loanword identification in low-resource language settings using a lexical-constrained GAN with two generators and a discriminator. It uses a log-linear RNN along with word and character-level embeddings, pronunciation similarity, and POS tagging features.

5.1.4 Fraud Detection

Social media platforms monitor user opinions on personal events, businesses, news, and politics. Market analysts use such reviews to come up with predictions and strategies to improve their business. To dominate the market, business owners may tend to add fake reviewers to their accounts or competitors' accounts. With the advancement of technology and bot usage, these fraud reviews are increasing exponentially. Hence, it is vital to identify such fraudulent reviews to perform a more reliable market analysis. There are different types of attempts in current research targeting fraud text detection, such as language models (Ott et al., 2011), behavioural profile analysis (Rayana and Akoglu, 2015) and deep learning feature representations (Le and Mikolov, 2014). A vital issue in fraud review identification is the lack of trusted labelled data, which leads to data scarcity of the models. To handle this problem, Aghakhani et al. (2018) proposed FakeGAN with one generator and two discriminators that address the model collapse problem, which is a typical problem for the GAN models. The training dataset X combines the subsets, X_T and X_D, which are fraud and real reviews, respectively. Z_g indicates all the reviews generated by FakeGAN. One discriminator, D, is defined for classifying fake ($X_D \cup Z_G$) and real X_T samples. Another discriminator, D′, is defined for classifying the generated samples similar to X_T and X_D. The model training follows the stochastic policy gradient method in reinforcement learning. Figure 5.5 illustrates an overview of FakeGAN, where the positive and negative samples are indicated by + and - symbols, respectively. The evaluation results of Aghakhani et al. (2018) indicate that the FakeGAN model performs similarly to the other fraud detection models in the literature. A main limitation of the model is the capability of generating reviews only in plain text without any association with the metadata, such as the rating scores. The possibility of bot-generated reviews in the training set as real samples and instability in the training process must also be addressed in future work. Further, another future research mentioned is the exploration of other GAN variants, such as Conditional GAN, and performing experiments with better hyperparameter tuning (Aghakhani et al., 2018).

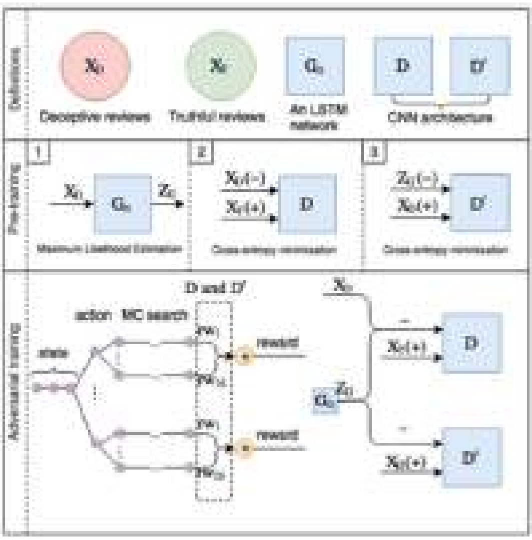

Figure 5.5: The overview of FakeGAN (Aghakhani et al., 2018)

The work proposed by Shehnepoor et al. (2022) addresses the drawbacks mentioned above of FakeGAN (Aghakhani et al., 2018) by generating score-correlated reviews using Information Gain Maximisation (IGM) theory to filter the fake samples that are generated. Their proposed model is called ScoreGAN, and it incorporates a given set of real reviews X, genuine reviews with scores $<X_g,S>$, fraud-human reviews with scores $<X_{fh},S>$ to generate score-correlated fraud bot reviews $<X_{fg},S>$. The overall fraud review set is $X_f = \{X_{fh},X_{fg}\}$. This model utilises two discriminators, D_g and D_f, following the FakeGAN architecture. The augmented data enables the discriminator D_g to distinguish bot-generated fraud reviews effectively. Figure 5.6 illustrates the framework of the ScoreGAN model. The information gain between the constraint c and the generator G_θ *(z,c)* is as follows:

$$I(c, G_\theta(z,c)) = H(c|G_\theta(z,c)) = -\mathbb{E}_{x\sim G_\theta(z,c), c\sim P(c|x)}[-\log P(c|x)] + H(c) \tag{14}$$

Using Lemma to address the issue of a fixed distribution on c, where H is the entropy definition, yields:

$$\begin{aligned} L(G_\theta, Q) &= -\mathbb{E}_{x\sim G_\theta(z,c), c\sim P(c|x)}[-\log Q(c|x)] + H(c) \\ &= \mathbb{E}_{x\sim G_\theta(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\leq I(c, G_\theta(z,c)) \end{aligned} \tag{15}$$

The overall minimax game for is defined as follows:

$$max(\mathbb{E}_{x\sim X_g}[\log D_g(x)] + \mathbb{E}_{x\sim X_{fh}}[1 - \log D_g(x)] + \lambda L(G_\theta, Q)) \tag{16}$$

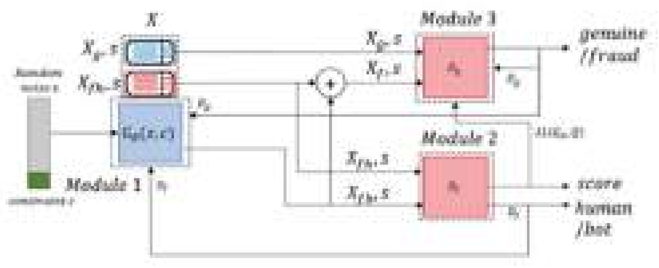

Figure 5.6: The illustration of the ScoreGAN model (Shehnepoor et al., 2022)

The evaluation results presented by Shehnepoor et al. (2022) showcase a 5% accuracy increase in Trip Advisor reviews and a 7% accuracy increase in Yelp reviews. Interestingly, experiments with a smaller subset of training data combined with augmented data are as effective as the full-sized datasets. A future direction in ScoreGAN would be to combine text features with other features, such as metadata (Shehnepoor et al., 2022).

Besides generating fraudulent reviews, social bots manipulate public opinions on different topics, accounts, and topics and spread malicious content. Due to the negative impacts that social bots impose, detecting and removing such fake accounts from social networks is nowadays crucial., It may lead to even more severe issues when the data generated by bots are more than those generated by genuine accounts because of the class imbalance issue. Wu et al. (2020) introduce an improved conditional GAN with a modified Gaussian Kernel Density Peak Clustering Algorithm (GKDPCA) to reduce noisy data generation and eliminate class imbalance within the data. The social bot detection framework uses a set of features: user-based, content and network. The use of Wasserstein distance with gradient penalty addresses the original conditional GAN model issues, which involve model collapse and the inability to control the category information in generated samples. As per the evaluation results, the improved cGAN outperforms three standard oversampling methods: random sampling (Liu et al., 2007), ADASYN (He et al., 2008) and SMOTE (Chawla et al., 2002) with a 97.56% of F1 score. As Wu et al. (2020) suggest, future work may head toward malicious bot detection incorporating other behavioural patterns and feature sequences.

Apart from the above applications, GAN text data augmentation has been employed for phishing URL detection to synthesise the training data (Xiao et al., 2021; Lee et al., 2020; Anand et al., 2018). Stanton and Irissappane (2019) present spamGAN for opinion spam detection that employs a semi-supervised GAN model.

5.1.5 Code-Switching Sentence Generation

Code-switching corresponds to the language changes in a given text. It may exist at the word or subword level when the editor writes different pieces in a text by changing it from one language to another. Chang et al. (2019) present an unsupervised GAN architecture to generate code-switching intra-sentences from monolingual data. Approaches to code-switching applications involve expensive human annotations and labelling speech data via transcription. (Chang et al., 2019) present a mechanism to generate such code-switching data without using any labelled data in the generator. Another application of GAN-based augmentation for code-switching is proposed by Gao et al. (2019) to generate intra-sentential code-switching sentences based on monolingual data, which outperforms code-switching language models. The future direction of Gao et al. (2019) will be towards enhancing the translator and generator.

5.1.6 Miscellaneous Applications

Large labelled dataset construction is a time-consuming process and requires domain expertise. Generative models with data augmentation are usually more sensitive to generating such categorically labelled data than complex manual annotation approaches. Most sentence generation models using GANs involve unlabelled texts, but it is also required to generate labelled data for a supervised classification task. There are two possible ways to perform this task: adding category information to the model or making the model generate a categorical sentence. The first approach loads the label information into the input representation. CS-GAN (Li et al., 2018) uses reinforcement learning, RNN and GAN-based category sentence generation to enlarge the original dataset. The sentiment analysis model by Li et al. (2018) performs well in supervised learning and shows the best performance with varying sentence lengths, even with smaller datasets with more categories.

Several other notable GAN application domains in text data augmentation include literary texts (Shahriar, 2022), multimodal news domain (Cadigan et al., 2021), controlled text generation (Betti et al., 2020; Malandrakis et al., 2019), machine translation (Ma et al., 2022; Fadaee et al., 2017; Sennrich et al., 2016) and medical domain (Kasthurirathne et al., 2021; Guan et al., 2018). These models either use GAN-synthesised data to mix with training data in pre-training or directly use the data generation alongside the training.

5.2 Critical Analysis of the Literature

Table 1 illustrates several applications of GAN text data augmentation in recent research in areas such as sentiment analysis, low resource language generation, fraud detection, code-switching sentence generation, and medical text generation, with a summary of approaches and future directions. Most models use SeqGAN architecture (Yu et al., 2017) with a few modifications in optimising the loss function. In category or label-based training, SentiGAN models Wang and Wan (2018) are adopted by providing label information and input features. Some of the models employ multiple generators or multiple discriminator architectures as well. Although not directly supporting text generation, Xiao et al. (2021) use Vanilla GAN to generate data GAN synthesised URLs. Future researchers could investigate enhancing these applications with a better combination of various features, enhancing training stability, extending to other languages, and building different GAN architectures.

Application	GAN Architecture	Approach	Suggested Future Directions
Sentiment Analysis	C-GAN (Gupta 2019)	Conditional GAN to augment data for sentiment classification with a generator, a discriminator, and a baseline classifier	Apply other GAN variants
	Seq-GAN (Luo et al., 2021)	Penalty-based SeqGAN to generate high-quality synthesised data	Use framework for other text domains
	G2S-AT-GAN (Chen et al., 2021)	Knowledge-graph-based rumour data augmentation (GERDA) and attention-based graph convolutions network with GAN	Address the problem of rumour data imbalance
	TransGAN (Shang et al., 2021)	RoBERTa model enhanced by a transformer-based GAN	Test the applicability of other datasets and cross-domain adaptation
Code-Switching Sentence Generation	Unsupervised GAN (Chang et al., 2019)	Unsupervised method to generate intra-sentential code-switching sentences using GAN	Improve translation accuracy
	CS-GAN (Gao et al., 2019)	Bert-C-based generator and discriminator	Generate a longer sequence of foreign words
Low-Resource Language Generation	QuGAN (Sun et al., 2019)	Tibetan question-answering corpus generation combining QuasiRNN and GAN	Increase the accuracy in generated corpus and add argument function and Tibetan grammar function
	Senti-GAN (Carrasco et al., 2021)	Sentimental GAN to generate sentences to overcome the data scarcity of the annotated Arabic regional dialects	Generate country-level dialects with data augmentation
	Lexical Controlled GAN (Mi et al., 2021)	Lexical constraint-based GAN to generate loanwords	Improve robustness of loanword identification with data augmentation

Table 1: Summary of GAN Text Augmentation Approaches

Application	**GAN Architecture**	**Approach**	**Suggested Future Directions**
Fraud Detection	Fake-GAN (Aghakhani et al., 2018)	Use two discriminator models and one generative model	Comparison with state-of-the-art supervised techniques
	Vanilla GAN (Xiao et al., 2021)	Use GAN-synthesised URLs to balance the datasets of legitimate and phishing URL	Explore the evolution pattern of the phishing websites
	Phish-GAN (Lee et al., 2020)	Use GAN to generate images of hieroglyphs conditioned on non-homoglyph input text images	Extend to other languages, such as Chinese and Korean
	C-GAN (Wu et al., 2020)	Improve the CGAN convergence issue by Wasserstein distance with a gradient penalty	Focus on malicious social bot detection
	Semi-Supervised GAN (Fadhel and Nyarko 2019)	Semi-supervised adversarial learning with discrete elements	Analysing the performance when incorporating the Movers distance measure
	Score-GAN (Shehnepoor et al., 2022)	Incorporate scores through IGM into the loss function	Combine text features with other behavioural features
Medical Text Generation	Seq-GAN (Kasthurirathne et al., 2021)	Generate synthetic free-text medical data with limited reidentification risk	
	mtGAN (Guan et al., 2018)	Generate synthetic texts of EMRs using reinforcement learning-based GAN	Explore hidden representations of medical texts

Table 1 (Continued): Summary of GAN Text Augmentation Approaches

6. Current Challenges and Future Research

The systematic review of GAN-based text data augmentation presented in this paper shows that many proposed frameworks for GAN-based text data augmentation still suffer from a lower accuracy for the classification tasks and the generation of grammatically incorrect long-textual data (Luo et al., 2021).

Evaluating the quality of the generated data is another potential gap in current research since there is a relatively lower number of attempts focusing on text data augmentation. There is still room for research on why and how data augmentation techniques provide accuracy improvements with a notion of in-depth theories and principles. In semantic classification methodologies involving data augmentation, it will be interesting to observe the impact of fake data generated on the opposition class via GANs to observe whether it will improve the model accuracy.

7. Conclusion

The paper provides a background study to showcase the recent research on GAN models as a text data augmentation tool. We used the PRISMA framework to ensure a non-biased and efficient paper search. With the notion of academic aspirations around data augmentation and GAN models, the paper presents a close view of applications spanning from sentence generation, addressing low resource languages, sentiment analysis and text analysis. Future directions in this area will further explore generating data distribution similar to but different from the original to reduce overfitting scenarios and new metrics to evaluate such text generation.

References

Aghakhani, H., Machiry, A., Nilizadeh, S., Krügel, C. and Vigna, G. (2018). Detecting Deceptive Reviews Using Generative Adversarial Networks. In: *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE Computer Society, 89–95. Available at: https://doi.org/10.1109%2Fspw.2018.00022.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N. and Zwerdling, N. (2020). Do Not Have Enough Data? Deep Learning to the Rescue!. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York: AAAI Press, 7383–7390. Available at: https://ojs.aaai.org/index. php/AAAI/article/view/6233.

Anand, A., Gorde, K., Antony Moniz, J. R., Park, N., Chakraborty, T. and Chu, B.-T. (2018). Phishing URL Detection with Oversampling based on Text Generative Adversarial Networks. In: Abe, N. et al., (eds.). *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1168–1177. Available at: https:// doi.org/10.1109%2Fbigdata.2018.8622547.

Arjovsky, M. and Bottou, L. (2017). Towards Principled Methods for Training Generative Adversarial Networks. In: *5th International Conference on Learning Rep-*

resentations, ICLR 2017. OpenReview.net. Available at: https://openreview.net/forum?id=Hk4_qw5xe.

Betti, F., Ramponi, G. and Piccardi, M. (2020). Controlled Text Generation with Adversarial Learning. In: Davis, B. et al., (eds.). *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*. Association for Computational Linguistics, 29–34. Available at: https://aclanthology.org/2020.inlg-1.5/.

Cadigan, J., Sikka, K., Ye, M. and Graciarena, M. (2021). Resilient Data Augmentation Approaches to Multimodal Verification in the News Domain. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Cao, R. and Lee, R. K.-W. (2020). HateGAN: Adversarial Generative-Based Data Augmentation for Hate Speech Detection. In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 6327–6338. Available at: https://aclanthology.org/2020.coling-main.557.

Carrasco, X. A., Elnagar, A. and Lataifeh, M. (2021). A Generative Adversarial Network for Data Augmentation: The Case of Arabic Regional Dialects. In: *Fifth International Conference On Arabic Computational Linguistics, ACLING 2021*. Online: Elsevier, 92–99. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921011674.

Chang, C.-T., Chuang, S.-P. and Lee, H. (2019). Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation. In: Kubin, G. and Kacic, Z., (eds.). *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 554–558. Available at: https://doi.org/10.21437%2Finterspeech.2019-3214.

Chawla, N. v, Bowyer, K. W., Hall, L. O. and Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. Available at: https://doi.org/10.1613%2Fjair.953.

Chen, H., Ji, Y. and Evans, D. (2020). Finding Friends and Flipping Frenemies: Automatic Paraphrase Dataset Augmentation Using Graph Theory. In: Cohn, T., He, Y., and Liu, Y., (eds.). *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 4741–4751. Available at: https://aclanthology.org/2020.findings-emnlp.426.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P. (2016). Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In: Lee, D. D. et al., (eds.). *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc., 2172–2180. Available at: https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.

Chen, X., Zhu, D., Lin, D. and Cao, D. (2021). Rumor Knowledge Embedding Based Data Augmentation for Imbalanced Rumor Detection. *Information Sciences,* 580, 352–370. Available at: https://doi.org/10.1016/j.ins.2021.08.059.

Donahue, D. and Rumshisky, A. (2018). Adversarial Text Generation Without Reinforcement Learning. *CoRR*, abs/1810.06640. Available at: http://arxiv.org/abs/1810.06640.

Fabbri, A., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D. and Mehdad, Y. (2021). Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In: Toutanova, K. et al., (eds.). *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 704–717. Available at: https://aclanthology.org/2021.naacl-main.57.

Fadaee, M., Bisazza, A. and Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. In: Barzilay, R. and Kan, M.-Y., (eds.). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 567–573. Available at: https://aclanthology.org/P17-2090.

Fadhel, M. ben and Nyarko, K. (2019). GAN Augmented Text Anomaly Detection with Sequences of Deep Statistics. In: *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 1–5. Available at: https://doi.org/10.1109/CISS.2019.8693024.

Fedus, W., Goodfellow, I. J. and Dai, A. M. (2018). MaskGAN: Better Text Generation via Filling in the ______. In: *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*. Available at: https://openreview.net/pdf?id=ByOExmWAb.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T. and Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. In: Zong, C. et al., (eds.). *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 968–988. Available at: https://aclanthology.org/2021.findings-acl.84.

Feng, S. Y., Gangal, V., Kang, D., Mitamura, T. and Hovy, E. (2020). GenAug: Data Augmentation for Finetuning Text Generators. In: *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Online: Association for Computational Linguistics, 29–42. Available at: https://aclanthology.org/2020.deelio-1.4.

Gao, Y., Feng, J., Liu, Y., Hou, L., Pan, X. and Ma, Y. (2019). Code-Switching Sentence Generation by Bert and Generative Adversarial Networks. In: Kubin, G. and Kacic, Z., (eds.). *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 3525–3529. Available at: https://doi.org/10.21437%2Finterspeech.2019-2501.

Goldberg, A. and Zhu, X. (2006). Seeing Stars When There Aren't Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. In: *Proceedings of TextGraphs: The First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 45–52. Available at: https://aclanthology.org/W06-3808.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y. (2014). Generative Adversarial Nets. In: Ghahramani, Z. et al., (eds.). *Advances in Neural Information Processing Systems 27: Annual Conference on NIPS 2014*, 2672–2680. Available at: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Guan, J., Li, R., Yu, S. and Zhang, X. (2018). Generation of Synthetic Electronic Medical Record Text. In: Zheng, H. J. et al., (eds.). *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE Computer Society, 374–380. Available at: https://doi.org/10.1109%2Fbibm.2018.8621223.

Guo, H. (2020). Nonlinear Mixup: Out-Of-Manifold Data Augmentation for Text Classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 4044–4051. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/5822.

Gupta, R. (2019). Data Augmentation for Low Resource Sentiment Analysis Using Generative Adversarial Networks. In: *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7380–7384. Available at: https://doi.org/10.1109%2Ficassp.2019.8682544.

Gupta, R., Sahu, S., Espy-Wilson, C. Y. and Narayanan, S. S. (2018). Semi-Supervised and Transfer Learning Approaches for Low Resource Sentiment Classification. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5109–5113. Available at: https://doi.org/10.1109/ICASSP.2018.8461414.

He, H., Bai, Y., Garcia, E. A. and Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 1322–1328. Available at: https://doi.org/10.1109/IJCNN.2008.4633969.

Hou, Y., Liu, Y., Che, W. and Liu, T. (2018). Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding. In: Bender, E. M., Derczynski, L., and Isabelle, P., (eds.). *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*. Association for Computational Linguistics, 1234–1245. Available at: https://aclanthology.org/C18-1105.

Jia, R., Raghunathan, A., Göksel, K. and Liang, P. (2019). Certified Robustness to Adversarial Word Substitutions. In: Inui, K. et al., (eds.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 4129–4142. Available at: https://aclanthology.org/D19-1423.

Kang, D., Khot, T., Sabharwal, A. and Hovy, E. (2018). AdvEntuRe: Adversarial Training for Textual Entailment with Knowledge-Guided Examples. In: Gurevych, I. and Miyao, Y., (eds.). *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2418–2428. Available at: https://aclanthology.org/P18-1225.

Karras, T., Laine, S. and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation/IEEE, 4396–4405. Available at: https://doi.org/10.1109%2Fcvpr.2019.00453.

Kasthurirathne, S. N., Dexter, G. and Grannis, S. (2021). Generative Adversarial Networks for Creating Synthetic Free-Text Medical Data: A Proposal for Collaborative Research and Re-use of Machine Learning Models. In: *Proceedings – AMIA Joint Summits Translational Science*, 335–344.

Kim, H.-Y., Roh, Y.-H. and Kim, Y.-K. (2019). Data Augmentation by Data Noising for Open-vocabulary Slots in Spoken Language Understanding. In: Kar, S. et al., (eds.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 97–102. Available at: https://aclanthology.org/N19-3014.

Kumar, A., Bhattamishra, S., Bhandari, M. and Talukdar, P. (2019). Submodular Optimization-based Diverse Paraphrasing and its Effectiveness in Data Augmentation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3609–3619. Available at: https://aclanthology.org/N19-1363.

Kusner, M. J. and Hernández-Lobato, J. M. (2016). GANS for Sequences of Discrete Elements with the Gumbel-softmax Distribution. *CoRR*, abs/1611.04051. Available at: http://arxiv.org/abs/1611.04051.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning – Volume 32*. PMLR, 1188–1196. Available at: http://proceedings.mlr.press/v32/le14.html.

Lee, J. S., Yam, G. P. D. and Chan, J. H. (2020). PhishGAN: Data Augmentation and Identification of Homoglpyh Attacks. In: *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 1–6. Available at: https://doi.org/10.1109%2Fccci49893.2020.9256804.

Li, Y., Pan, Q., Wang, S., Yang, T. and Cambria, E. (2018). A Generative Model for Category Text Generation. *Information Sciences*, 450, 301–315. Available at: https://doi.org/10.1016%2Fj.ins.2018.03.050.

Liu, A. Y., Ghosh, J. and Martin, C. E. (2007). Generative Oversampling for Mining Imbalanced Datasets. In: Stahlbock, R., Crone, S. F., and Lessmann, S., (eds.).

Proceedings of the 2007 International Conference on Data Mining, DMIN. CSREA Press, 66–72.

Liu, M.-Y. and Tuzel, O. (2016). Coupled Generative Adversarial Networks. In: Lee, D. D. et al., (eds.). *Advances in Neural Information Processing Systems 29: Annual Conference on NIPS 2016*. NeurIPS, 469–477. Available at: https://proceedings.neurips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html.

Longpre, S., Lu, Y., Tu, Z. and DuBois, C. (2019). An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 220–227. Available at: https://aclanthology.org/D19-5829.

Louvan, S. and Magnini, B. (2020). Simple is Better! Lightweight Data Augmentation for Low Resource Slot Filling and Intent Classification. In: Nguyen, M. le, Luong, M. C., and Song, S., (eds.). *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics, 167–177. Available at: https://aclanthology.org/2020.paclic-1.20.

Lu, K., Mardziel, P., Wu, F., Amancharla, P. and Datta, A. (2020). Gender Bias in Neural Natural Language Processing. In: Nigam, V. et al., (eds.). *Logic, Language, and Security – Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*. Springer, 189–202. Available at: https://doi.org/10.1007/978-3-030-62077-6_14.

Luo, J., Bouazizi, M. and Ohtsuki, T. (2021). Data Augmentation for Sentiment Analysis Using Sentence Compression-Based SeqGAN With Data Screening. *IEEE*, 9, 99922–99931. Available at: https://doi.org/10.1109%2Faccess.2021.3094023.

Ma, W., Yan, B. and Sun, L. (2022). Generative Adversarial Network-based Short Sequence Machine Translation from Chinese to English. *Scientific Programming*, 2022, 1–10. Available at: https://doi.org/10.1155%2F2022%2F7700467.

Malandrakis, N., Shen, M., Goyal, A., Gao, S., Sethi, A. and Metallinou, A. (2019). Controlled Text Generation for Data Augmentation in Intelligent Artificial Agents. In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, 90–98. Available at: https://aclanthology.org/D19-5609.

Mao, X., Wang, Y., Liu, X. and Guo, Y. (2017). An Adaptive Weighted Least Square Support Vector Regression for Hysteresis in Piezoelectric Actuators. *Sensors and Actuators A: Physical*, 263, 423–429. Available at: https://doi.org/10.1016%2Fj.sna.2017.06.030.

Mi, C., Zhu, S. and Nie, R. (2021). Improving Loanword Identification in Low-Resource Language with Data Augmentation and Multiple Feature Fusion. *Computational Intelligence and Neuroscience*, 2021, 1–9. Available at: https://doi.org/10.1155%2F2021%2F9975078.

Mimura, M. (2020). Using Fake Text Vectors to Improve the Sensitivity of Minority Class for Macro Malware Detection. *Journal of Information Security and Ap-*

plications, 54, 102600. Available at: https://www.sciencedirect.com/science/article/pii/S2214212620307651.

Mirza, M. and Osindero, S. (2014). Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784. Available at: http://arxiv.org/abs/1411.1784.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-analyses: the PRISMA statement. *BMJ*, 339, b2535–b2535. Available at: https://www.bmj.com/content/339/bmj.b2535.

Nowozin, S., Cseke, B. and Tomioka, R. (2016). f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In: Lee, D. D. et al., (eds.). *Advances in Neural Information Processing Systems 29: Annual Conference on NIPS 2016*. NeurIPS, 271–279. Available at: https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html.

Ott, M., Choi, Y., Cardie, C. and Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In: Lin, D., Matsumoto, Y., and Mihalcea, R., (eds.). *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 309–319. Available at: https://aclanthology.org/P11-1032.

Parida, S. and Motlicek, P. (2019). Abstract Text Summarization: A Low Resource Challenge. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 5994–5998. Available at: https://aclanthology.org/D19-1616.

Park, T., Liu, M.-Y., Wang, T.-C. and Zhu, J.-Y. (2019). GauGAN: Semantic Image Synthesis with Spatially Adaptive Normalization. In: *ACM SIGGRAPH 2019 Real-Time Live!* Association for Computing Machinery. Available at: https://doi.org/10.1145%2F3306305.3332370.

Paschali, M., Simson, W., Roy, A. G., Naeem, M.F., Göbl, R., Wachinger, C. and Navab, N. (2019). Manifold Exploring Data Augmentation with Geometric Transformations for Increased Performance and Robustness. In: Chung Albert C. S. and Gee, J. C. and Y. P. A. and B. S., (eds.). *Information Processing in Medical Imaging.* Cham: Springer International Publishing, 517–529. Available at: https://doi.org/10.1007%2F978-3-030-20351-1_40.

Qin, L., Ni, M., Zhang, Y. and Che, W. (2021). CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. In: Bessiere, C., (ed.). *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI'20. International Joint Conferences on Artificial Intelligence Organization, 3853–3860. Available at: https://doi.org/10.24963/ijcai.2020/533.

Quan, J. and Xiong, D. (2019). Effective Data Augmentation Approaches to End-to-End Task-Oriented Dialogue. In: *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 47–52. Available at: https://doi.org/10.1109%2Fialp48816.2019.9037690.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8), 9.

Radford, A., Metz, L. and Chintala, S. (2018). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In: *2018 37th Chinese Control Conference (CCC)*. IEEE, 9159–9163. Available at: https://doi.org/10.23919%2Fchicc.2018.8482813.

Rayana, S. and Akoglu, L. (2015). Collective Opinion Spam Detection: Bridging Review Networks and Metadata. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery, 985–994. Available at: https://doi.org/10.1145/2783258.2783370.

Riabi, A., Scialom, T., Keraron, R., Sagot, B., Seddah, D. and Staiano, J. (2021). Synthetic Data Augmentation for Zero-Shot Cross-Lingual Question Answering. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7016–7030. Available at: https://aclanthology.org/2021.emnlp-main.562.

Şahin, G.G. and Steedman, M. (2018). Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5004–5009. Available at: https://aclanthology.org/D18-1545.

Sennrich, R., Haddow, B. and Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin: Association for Computational Linguistics, 86–96. Available at: https://aclanthology.org/P16-1009.

Shahriar, S. (2022). GAN Computers Generate Arts? A Survey on Visual Arts, Music, and Literary Text Generation using Generative Adversarial Network. *Displays*, 73, 102237. Available at: https://www.sciencedirect.com/science/article/pii/S0141938222000658.

Shang, Y., Su, X., Xiao, Z. and Chen, Z. (2021). Campus Sentiment Analysis with GAN-based Data Augmentation. In: *13th International Conference on Advanced Infocomm Technology (ICAIT)*. IEEE, 209–214. Available at: https://doi.org/10.1109%2Ficait52638.2021.9702068.

Shehnepoor, S., Togneri, R., Liu, W. and Bennamoun, M. (2022). ScoreGAN: A Fraud Review Detector Based on Regulated GAN With Data Augmentation. *IEEE Transactions on Information Forensics and Security*, 17, 280–291.

Sindhwani, V. and Melville, P. (2008). Document-Word Co-regularization for Semi-supervised Sentiment Analysis. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE Computer Society, 1025–1030. Available at: https://doi.org/10.1109/ICDM.2008.113.

Socher, R., Pennington, J., Huang, E. H.-C., Ng, A. and Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions.

In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 151–161. Available at: https://aclanthology.org/D11-1014/.

Stanton, G. and Irissappane, A. A. (2019). GANs for Semi-Supervised Opinion Spam Detection. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 5204–5210. Available at: https://doi.org/10.24963/ijcai.2019/723.

Sun, Y., Chen, C., Xia, T. and Zhao, X. (2019). QuGAN: Quasi Generative Adversarial Network for Tibetan Question Answering Corpus Generation. *IEEE Access*, 7, 116247–116255. Available at: https://doi.org/10.1109%2Faccess.2019.2934581.

Täckström, O. and McDonald, R. T. (2011). Semi-supervised Latent Variable Models for Sentence-level Sentiment Analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 569–574. Available at: https://aclanthology.org/P11-2100.

Wang, K. and Wan, X. (2018). SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 4446–4452. Available at: https://doi.org/10.24963%2Fijcai.2018%2F618.

Wei, J. and Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 6382–6388. Available at: https://aclanthology.org/D19-1670.

Wu, B., Liu, L., Yang, Y., Zheng, K. and Wang, X. (2020). Using Improved Conditional Generative Adversarial Networks to Detect Social Bots on Twitter. *IEEE Access*, 8, 36664–36680. Available at: https://doi.org/10.1109%2Faccess.2020.2975630.

Wu, J., Zhang, C., Xue, T., Freeman, B. and Tenenbaum, J. (2016). Learning a Probabilistic Latent Space of Object Shapes via 3d Generative-adversarial Modeling. In: Lee, D. D. et al., (eds.). *Advances in neural information processing systems*. Barcelona, 82–90. Available at: https://proceedings.neurips.cc/paper/2016/hash/44f683a84163b3523afe57c2e008bc8c-Abstract.html.

Xiao, X., Xiao, W., Zhang, D., Zhang, B., Hu, G., Li, Q. and Xia, S. (2021). Phishing Websites Detection via CNN and Multi-head Self-attention on Imbalanced Datasets. *Computers & Security*, 108, 102372. Available at: https://www.sciencedirect.com/science/article/pii/S0167404821001966.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T. and Le, Q. v. (2020). Unsupervised Data Augmentation for Consistency Training. In: Larochelle, H. et al., (eds.). *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Red Hook: Curran Associates Inc., 6256–6268. Available at: https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M. and Lin, J. J. (2019). Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering. *CoRR*, abs/1904.06652. Available at: https://arxiv.org/abs/1904.06652.

Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., le Bras, R., Wang, J.-P., Bhagavatula, C., Choi, Y. and Downey, D. (2020). Generative Data Augmentation for Commonsense Reasoning. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 1008–1025. Available at: https://aclanthology.org/2020.findings-emnlp.90.

Yu, L., Zhang, W., Wang, J. and Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In: Singh, S. and Markovitch, S., (eds.). *AAAI Conference on Artificial Intelligence*. AAAI Press, 2852–2858. Available at: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14344.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J. and Yoo, Y. J. (2019). CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 6022–6031. Available at: https://doi.org/10.1109%2Ficcv.2019.00612.

Zhang, H., Cissé, M., Dauphin, Y. and Lopez-Paz, D. (2018). Mixup: Beyond Empirical Risk Minimization. In: *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net. Available at: https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.N. (2017). StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 5908–5916. Available at: https://doi.org/10.1109%2Ficcv.2017.629.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 15–20. Available at: https://aclanthology.org/N18-2003.

Zhu, H., Dong, L., Wei, F., Qin, B. and Liu, T. (2022). Transforming Wikipedia into Augmented Data for Query-Focused Summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2357–2367. Available at: https://doi.org/10.1109%2Ftaslp.2022.3171963.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2242–2251. Available at: https://doi.org/10.1109%2Ficcv.2017.244.