Text Generation to Aid Depression Detection: A Comparative Study of Conditional Sequence Generative Adversarial Networks

ML Tlachac, Walter Gerych, Kratika Agrawal, Benjamin Litterer, Nicholas Jurovich, Saitheeraj Thatigotla, Jidapa Thadajarassiri, and Elke A. Rundensteiner

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020224

***Abstract*—Corpuses of unstructured textual data, such as text messages between individuals, are often predictive of medical issues such as depression. The text data usually used in healthcare applications has high value and great variety, but is typically small in volume. Generating labeled unstructured text data is important to improve models by augmenting these small datasets, as well as to facilitate anonymization. While methods for labeled data generation exist, not all of them generalize well to small datasets. In this work, we thus perform a much needed systematic comparison of conditional text generation models that are promising for small datasets due to their unified architectures. We identify and implement a family of nine conditional sequence generative adversarial networks for text generation, which we collectively refer to as cSeqGAN models. These models are characterized along two orthogonal design dimensions: weighting strategies and feedback mechanisms. We conduct a comparative study evaluating the generation ability of the nine cSeqGAN models on three diverse text datasets with depression and sentiment labels. To assess the quality and realism of the generated text, we use standard machine learning metrics as well as human assessment via a user study. While the unconditioned models produced predictive text, the cSeqGAN models produced more realistic text. Our comparative study lays a solid foundation and provides important insights for further text generation research, particularly for the small datasets common within the healthcare domain.**

***Index Terms*—natural language processing, text classification, sentiment detection, digital phenotype, transfer learning**

I. Introduction

A. Motivation.

With the growing global mental health crisis [1], digital phenotype data – moment-by-moment quantification of individuals' behavior using data from personal digital devices [2] – is being explored as a strategy to make mental illness screening less burdensome and more universal. In particular, unstructured text data such as text messages [3] and transcripts [4] have demonstrated usefulness in screening for depression.

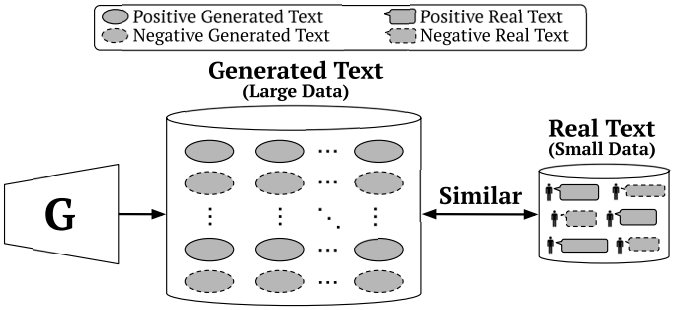


Fig. 1: A conditional text generation model inputs real labeled data instances to generate similar new labeled data instances. This is particularly important for healthcare applications like depression detection where the datasets tend to be small.

Unfortunately, it is challenging to collect large labeled datasets in certain domains such as healthcare where datasets must be carefully anonymized [5], [6]. Consequently, existing publicly available digital phenotype datasets only have between 48 and 369 participants [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. At the same time, unstructured text exhibits great variety. For instance, different individuals may have very different styles when composing text messages. The small volume combined with high variety is a serious issue, as data with more variety typically requires greater sample sizes to produce statistically significant results [18].

To advance critical diagnostic and prognostic modeling innovations, it is thus important to generate data realistic synthetic data [19]. In addition to increasing data quantity [20], data generation can help balance classes, amplify the class signal, and anonymize real data [21]. As *labeled* text data is required to train text classifiers, such healthcare applications would greatly benefit from *conditional* text generation so that class-specific data can be synthesized [22].

B. State-of-The-Art.

Generative adversarial networks (GANs) [23] revolutionised data generation with an architecture that involves a generator and discriminator. While the original GANs focused on images, text generation has received increased attention in recent years.

ML Tlachac is with Department of Information Systems & Analytics and Center for Health & Behavioral Sciences, Bryant University, Smithfield, RI 02911 USA, email: mltlachac@bryant.edu. Research completed while a PhD candidate at Worcester Polytechnic Institute (WPI).

Walter Gerych, Kratika Agrawal, Nicholas Jurovich, Jidapa Thadajarassiri, and Elke Rundensteiner are with Data Science and Computer Science Departments, Worcester Polytechnic Institute (WPI), Worcester, MA 01609 USA, emails: {wgerch,kagrawal,nsjurovich,jthadajarassiri,rundenst}@wpi.edu.

Benjamin Litterer is with the School of Information, University of Michigan, Ann Arbor, MI 48109 USA, email: blitt@umich.edu. Research completed while at WPI for NSF REU.

Saitheeraj Thatigotla is with the University of Tennessee, Knoxville, TN, 37996, email: sthatigo@vols.utk.edu. Research completed while at WPI for NSF REU.

Most approaches perform *unconditioned* text generation [24], [25], [26], [27], [28], [29], [30], where the generated data is unlabeled. Sequence GAN, known as SeqGAN [24], is a particularly popular approach.

When data quantities are large, a separate unconditioned generative model could be trained for each class. However, there are multiple advantages to *conditional* text generation where the generated text can be controlled to match specific classes. Notably, conditional models are trained with data from all classes which allows for the sharing of parameters, resulting in better language learning and applicability to smaller datasets.

The initial approaches for conditional text generation relied on *multiple generators* [31] or *discriminators* [32], which negates many benefits of conditional modeling by limiting parameter sharing. Also, these approaches do not scale well for use cases with many classes. There are a few other promising text generation methods [33], [6], [34] with a unified architecture containing only one generator and discriminator. However, some fundamental conditional approaches for images such as conditional GANs [22] have yet to be adapted and assessed for text generation.

C. Problem Statement.

A strategic assessment and comparative study of fundamental conditional text generation approaches is needed. The goal of conditional generation models, depicted in Figure 1, is to input a small labeled dataset to generate a large quantity of realistic labeled data with samples for each class. To work well for small datasets, such conditional generative models should have unified architectures that allow for parameter sharing across class labels during training. While there are a few conditional text generation approaches that leverage a unified architecture [33], [6], [34], there has been no identification nor categorization of fundamental conditional design approaches. Further, it is unknown which designs will be most effective for small datasets.

D. Our Approach.

We are the first to conduct an extensive study of unified text generative architectures, namely, those that have a single generator and a single discriminator. These unified architectures are not only appropriate for small healthcare datasets, but also easily scale for any number of classes, increasing their utility. By analyzing existing methods in the literature [22], [31], [32], [33], [6], [35], we identify two core orthogonal design dimensions compatible with unified sequence generation architectures; namely, three weighting strategies and three feedback mechanisms.

To tackle the problem of assessing different conditional text generative approaches, we thus compose each of the alternate weighting strategies and feedback mechanisms into a total of nine fundamental conditional text generation models. While implemented with the popular SeqGAN architecture, these approaches could be integrated with any recurrent generative adversarial model.

We leverage this family of nine models to generate text. In addition to two datasets with depression screening labels, we also demonstrate our methods on a publicly available dataset in the related domain of sentiment detection [36]. These datasets exemplify the *variety* and *value* emblematic of big data [37]. The variety comes from the fact that the text in each corpus was generated by many different individuals, each with a unique style of writing. The value comes from the depression labels associated with two of the datasets; passively detecting depression from text would provide immense value towards automatic screening systems [38].

We perform a comprehensive comparative study to evaluate the ability of the cSeqGAN and non-conditional SeqGAN models to generate realistic and predictive text. We evaluate the performance of all models through machine evaluation by using a pretrained BERT language model [39]. Additionally, we also designed a user study to obtain human assessment of the generated text.

E. Our Contributions

Contributions of our research include:

1) Identifying and adapting three weighting strategies and three feedback mechanisms for conditional text generation.
2) Assembling nine scalable cSeqGAN models that are applicable to small datasets that are common in healthcare.
3) Implementing of our models within a text generation benchmarking platform to assist future researchers.
4) Evaluating the nine cSeqGAN models on three real-world datasets with depression and sentiment labels.

II. Related Work in Generative Modeling

GANs [23] involve a minimax game between a generator and discriminator to continually improve the quality of generated text. Variational autoencoders (VAEs) [40] are also often used to generate data; an encoder maps the real data to a latent input space and a decoder uses this latent space to construct new data instances. As VAEs perform probabilistic sampling, the data is much more uniform than that generated by GANs. For many healthcare applications, the more realistic and diverse output of GANs is preferable. Conditional GANs [22] proposed the first conditional generative approach for images; the one hot encoded class labels were concatenated with the input to both the generator and discriminator.

SeqGAN [24] was first to adapt GANs for text. Notably, SeqGAN employed a sequential decision making process and a Monte Carlo search to approximate state-action value pairs [24]. These adaptations were vital to overcome the limited dictionary space inherent from using tokens rather than continuous values and reduce the sparsity of rewards.

Since then, there have been other proposed GANs for text generation: MaliGAN with Maximum Likelihood Augmentation [25], GSGAN which uses a Gumbel-softmax Distribution [26], RankGAN with adversarial ranking [27], TextGAN with adversarial feature matching [41], LeakGAN which improves long text generation with leaked information [29], and RelGAN with a relational memory based generator [30]. All but the last are encompassed in the Texygen benchmarking platform [42] which aims to standardize text

generation research; SeqGAN and LeakGAN performed best, motivating our use of SeqGAN in this research.

There have been a few recent attempts at conditional text generation. The first approach [32] involves introducing adversarial training to a VAE by incorporating a discriminator for each class option. Meanwhile, SentiGAN [31] proposes an architecture with a separate generator for each class option but a single discriminator. These approaches unfortunately do not scale well. While category sentence generative adversarial network (CS-GAN) [33] introduces an auxiliary classifier, only one classifier, one discriminator, and one generator are required regardless of the number of class options. Medical Text Generative Adversarial Network (mtGAN) [6] introduces a conditional constraint to SeqGAN by including features as additional input at every step of the sequence generation process. Most recently, category-aware GAN (CatGAN) [34] uses a hierarchical evolutionary learning algorithm to generate text in an approach that deviates far from the traditional SeqGAN architecture. Some of these text generation approaches are quite promising, but no strategic comparative study has been conducted to compare fundamental conditioning strategies.

III. Text Generation Methodology

We compare the three *weighting strategies* and three *feedback mechanisms* we identified for conditional text generation. The *weighting strategies* refer to design choices for how the previously generated word along with the class of the overall text determine the subsequently generated word. The *feedback mechanisms* refer to architectural choices for the critic that determines the realism of the data for the given class.

These two design dimensions are orthogonal, in that each of the possible choices along one dimension can be integrated with all the other choices of the second dimension. That is, a model can be designed to support one of the weighting strategies and one of the feedback mechanisms. In this work, we embed both types of strategies within the popular SeqGAN model architecture [24], thus constructing a total of nine unique conditional sequence generative adversarial networks (cSeqGAN) architectures. Given the unified architecture of these nine cSeqGAN models, we anticipate that they can be trained to generate text with relatively small datasets.

A. Sequence GANs

As with traditional GANs [23], SeqGANs [24] consist of a generator and a discriminator involved in a minimax game that iteratively improves text quality. However, SeqGAN alters the generation process to make it applicable for sequences of discrete tokens. For text generation, the tokens are words.

SeqGAN leverages a recurrent neural network (RNN) with Long Short Term Memory (LSTM) [43] as the generator. The LSTM maps each embedded word $x_t \in x_1, \ldots, x_T$ to a hidden state h_t to create a sequence of hidden states $h_1, \ldots, h_T$. Notably, the LSTM implements the update function g in

$$h_t = g(h_{t-1}, x_t) \tag{1}$$

to prevent the vanishing and exploding gradient problem. A softmax output layer then maps these hidden states into an output token distribution.

The SeqGAN generator [24] has the objective of maximizing an expected end reward given the starting state s_0. This expected end reward is calculated using an action-value function of a sequence. Specifically, the REINFORCE algorithm [44] is used to estimate the action-value function. Further, seqGAN [24] evaluates the intermediate state action-value pairs to provide more frequent rewards. This is accomplished using a Monte Carlo search with a roll-out policy that samples the remaining $T - t$ tokens in the sequence. The roll-out policy is repeated to reduce variance and improve the reward estimations.

As is the case with GAN [23], SeqGAN uses a convolutional neural network (CNN) [28] as the discriminator. This CNN is responsible for classifying each input text as real or fake. Specifically, a convolution operation is applied to the token embeddings to produce feature maps which are then pooled. The goal [24] is to minimize the sigmoid cross entropy loss.

B. Design Dimension: Weighting Strategies

We study three different weighting strategies to condition SeqGAN models for conditional text generation. In particular, we study *sentence weighting* which directly adapts conditional GANs [22] to be applicable for text instead of images, and two different unit weighting strategies inspired by mtGAN [6]. These latter strategies include *single unit weighting* in which the generative model is repeatedly conditioned when generating each word of the sentence, and *dual unit weighting* which likewise conditions the generation of each word but learns separate weights for the words and labels respectively. These three weighting strategies are applicable for any model with a recurrent network for a generator.

1) Sentence Weighting: In this first approach, we condition the generator only once while generating each sentence. The initial input to the generator is x_0, where x_0 is the concatenation of z and y. $z \in \mathbb{R}^n$ is a draw from a multivariate Gaussian distribution, and y is the class we aim to generate. The generator is an LSTM, such that $x_t = LSTM(x_{t-1})$. Thus, the t word in the sentence is simply the output of the LSTM conditioned on the previous output, such that the initial input contains information of the class to be generated. The Generator LSTM is defined as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{3}$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$
$$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{5}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{6}$$
$$h_t = o_t \circ \sigma_h(c_t), \tag{7}$$

where f_t, i_t, o_t, and $\tilde{c}_t$ are the forget gate, update gate, output gate, and cell input activation vectors respectively, h_t is the hidden state, and c_t is the cell state vector.

This means the LSTM generator is only conditioned on the class *once* and is then tasked with generating the entire sentence.

2) Single Unit Weighting: We modify the previous strategy by conditioning the LSTM on the desired class *at each step* of the generation process instead of once at the start. The Single Unit Weighting takes the form of $x_t = LSTM(W(x_{t-1} \oplus y))$, where W is a weight on the input and $\oplus$ is the concatenation operator. Notably, both the word embedding x_t and class embedding y share the same weight W.

This is done by replacing x_t in Equations 14-17 with $q = x_t + y$, resulting in an LSTM defined as

$$f_t = \sigma_g(W_f q_t + U_f h_{t-1} + b_f) \tag{8}$$
$$i_t = \sigma_g(W_i q_t + U_i h_{t-1} + b_i) \tag{9}$$
$$o_t = \sigma_g(W_o q_t + U_o h_{t-1} + b_o) \tag{10}$$
$$\tilde{c}_t = \sigma_c(W_c q_t + U_c h_{t-1} + b_c) \tag{11}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{12}$$
$$h_t = o_t \circ \sigma_h(c_t). \tag{13}$$

3) Dual Unit Weighting.: We study another weighting strategy in which the class embedding and word embedding are *separate* inputs into the LSTM generator. In this approach, the generator takes the form of $x_t = LSTM(Wx_{t-1} \oplus Vy)$. W and V are separate weight matrices for the word embedding x_{t-1} and class embedding y. This approach with separate weights is in contrast to the previous approach that used a single weight for both features and labels. The LSTM is specifically defined as

$$f_t = \sigma_g(W_f x_t + V_f y + U_f h_{t-1} + b_f) \tag{14}$$
$$i_t = \sigma_g(W_i x_t + V_i y + U_i h_{t-1} + b_i) \tag{15}$$
$$o_t = \sigma_g(W_o x_t + V_o y + U_o h_{t-1} + b_o) \tag{16}$$
$$\tilde{c}_t = \sigma_c(W_c x_t + V_c y + U_c h_{t-1} + b_c) \tag{17}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{18}$$
$$h_t = o_t \circ \sigma_h(c_t), \tag{19}$$

where the Ws and Vs are separate weight matrices for the word embedding x and class embedding y.

C. Design Dimension: Feedback Mechanisms

We study three feedback mechanisms to adapt SeqGAN models for conditional generation. These feedback mechanisms can coexist with any of the aforementioned weighting strategies. The first feedback mechanism, which we refer to as *single task feedback*, directly adapts the conditional GANs discriminator [22]. The other two feedback mechanisms involve two separate tasks: assessing realness and assessing class appropriateness. We refer to the mechanism that only uses one critic as *dual task feedback* and the mechanism that uses two critics as *dual critic feedback*. This latter strategy is modeled after the dual critics in CS-GAN for text generation [33] and GAN-control for image generation [35]. While we use a CNN discriminator like SeqGAN, these three feedback mechanisms are applicable for any generative adversarial model that uses a discriminator.

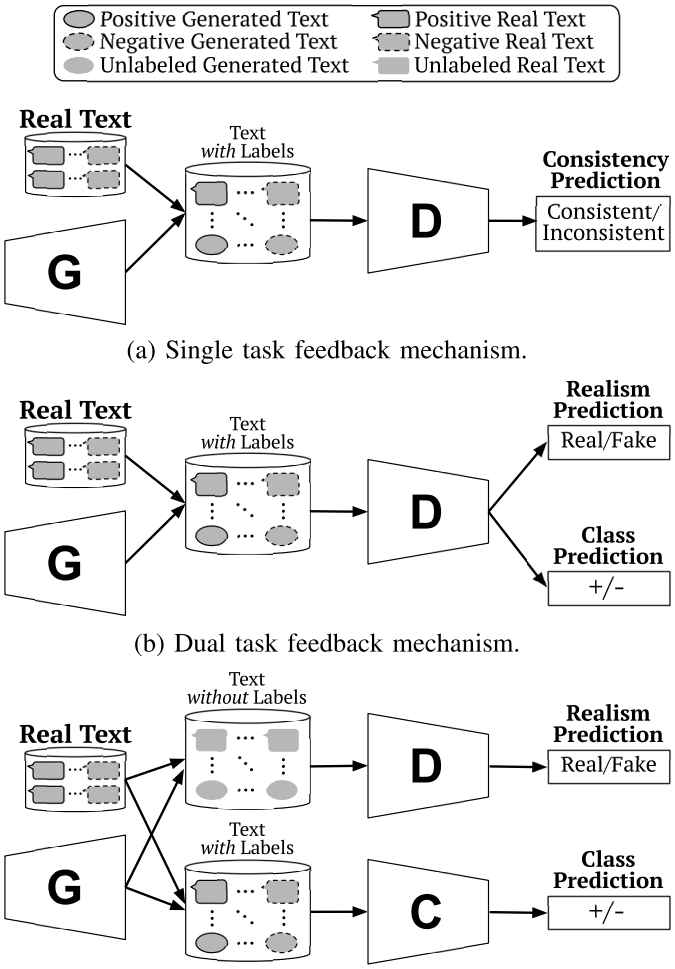
1) Single Task Feedback.: In this feedback mechanism, a single discriminator network is responsible for deciding whether the generated text is realistic *and* whether the generated text matches the condition y as a *single task*. Specifically, the discriminator D is a CNN that is trained with the following cross entropy loss function L_D:

$$L_D = \mathbb{E}_{d_g \sim (y,z)}[-log(1 - D(G(y,z),y))] + \mathbb{E}_{d_r \sim data}[log(D(d_r,y))] \tag{20}$$

Thus, D is trained to distinguish between real and generated data *given knowledge of the class label*, allowing for it to reject text that may otherwise be realistic but does not match the conditioning label. In this case, the generator G is trained with the loss L_G:

$$L_G = \mathbb{E}_{d_g \sim (y,z)}[log(D(G(y,z),y))] \tag{21}$$

2) Dual Task Feedback.: In this feedback mechanism, a single discriminator network performs two separate tasks: it determines whether the generated text is realistic and it determines whether the generated text matches the class label. Let $D_i^1(\tilde{x})$ be D's prediction probability for $y = i$, and let



(a) Single task feedback mechanism.

(b) Dual task feedback mechanism.

(c) Dual critic feedback mechanism.

Fig. 2: The three feedback mechanisms include: (a) a discriminator that performs a single task, (b) a discriminator that performs two tasks, and (c) a discriminator that performs a single task and a classifier that performs a single task. The realism prediction in (b) and (c) only consider the text while the consistency prediction in (a) considers the realness of the concatenated text and label.

$D^2(\tilde{x})$ be D's prediction for whether $\tilde{x}$ is real or generated. D is trained with the loss L_D:

$$\begin{aligned} L_D = &\mathbb{E}_{d_g \sim (y,z)}[-log(1 - D^2(G(y,z)))] \\ &+ \mathbb{E}_{d_r \sim data}[log(D^2(d_r))] \\ &+ \mathbb{E}_{(x,y) \sim data}[log(D^1_y(x))] \end{aligned} \tag{22}$$

In this case, the discriminator is trained to distinguish between real and generated text, while also being trained to correctly classify text. The generator is trained with L_G:

$$\begin{aligned} L_G = &\mathbb{E}_{d_g \sim (y,z)}[log(D^2(G(y,z),y))] \\ &+ \mathbb{E}_{(x,y) \sim (y,z)}[log(D^1_y(x))] \end{aligned} \tag{23}$$

In this setting the generator is trained to fool the discriminator into classifying it as real, while also being given the correct class by the discriminator.

3) Dual Critic Feedback.: This mechanism aims to separate the tasks of determining real text from fake text and correctly predicting classes. To this end, we utilize a *dual critic* approach in which the discriminator D only determines real text from fake text while a completely separate classifier C performs the classification. The discriminator D is thus trained as follows:

$$\begin{aligned} L_D = &\mathbb{E}_{d_g \sim (y,z)}[-log(1 - D(G(y,z)))] \\ &+ \mathbb{E}_{d_r \sim data}[log(D(d_r))] \end{aligned} \tag{24}$$

Thus, D is trained to determine real text from fake text with no information regarding the class. Meanwhile, C is trained to predict classes on the real data with loss L_C:

$$L_C = \mathbb{E}_{(x,y) \sim (y,z)}[log(C(x))] \tag{25}$$

Lastly, the generator G is trained to generate realistic enough text to fool D while achieving accurate classification of the generated text from C:

$$\begin{aligned} L_G = &\mathbb{E}_{d_g \sim (y,z)}[log(D(G(y,z)))] \\ &+ \mathbb{E}_{(x,y) \sim (y,z)}[log(C(x))] \end{aligned} \tag{26}$$

D. Combining Strategies

As mentioned, each of the three weighting strategies can be combined with each of three feedback mechanisms for conditional text generation. There are thus nine unique cSeqGAN architectures that can be constructed. Notably, not all combinations of the weighting strategies and feedback mechanisms exist in the related literature.

E. Software and Availability.

The code for our cSeqGAN models, implemented within the Texygen [42] framework, will be made available at *https://github.com/mltlachac/cSeqGAN* upon publication. Additionally, we will make the evaluation metrics for each individual model available. Due to sharing restrictions and privacy concerns, we are unable to share the raw transcripts or text messages. However, we have included examples of generated texts from our user study surveys in Tables IV - VI. While we focused our analysis on BLEU-2 due to the short lengths of the texts, we also include BLEU-3 and BLEU-4 scores of the generative models in Tables I - III. We will post research updates at *emutivo.wpi.edu*.

IV. Experimental Design

A. Datasets with Labeled Text

In this research, we leverage two datasets containing text with depression labels that have previously proved most useful for depression screening [4], [3]. Similar to many healthcare datasets [5], these datasets are small and would likely benefit from generative modeling to increase data quantity. Additionally, we demonstrate our methods on a popular publicly available dataset [45] in the related domain of sentiment detection [36] for replicability purposes.

Clinical Interviews. The Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset [46], [47] contains clinical interview transcripts labeled with Patient Health Questionnaire-8 (PHQ-8) depression screening scores. Available to academics upon request, these interviews are common for depression screening research [48]. A PHQ-8 score of at least 10 out of the possible 24 is interpreted as screening positive for depression [49]. Each of the 189 participants were asked a subset of core clinical interview questions by a virtual interviewer with followup questions as needed. For this research, we consider each of the 3774 sentences in the transcripts to be separate data instances.

Text Messages. The SMS text messages in the combined Moodable and EMU datasets [3], [9], [12] are labeled with PHQ-9 depression screening scores. While the PHQ-9 contains one more question than the PHQ-8, they share the same moderate depression screening cutoff of 10 [49]. These text messages were obtained from crowd-sourced study participants. Since the text messages capture real communications and are not in response to clinical prompts, we consider only texts from the participants with more polarizing PHQ-9 scores. Specifically, we use the 5360 text messages sent within the prior two weeks by the crowdsourced participants with PHQ-9≤ 5 and PHQ-9≥ 15.

Movie Reviews. The publicly available Stanford's Large Movie Review Dataset [45], commonly referred to as the Internet Movie Database (IMDb) Movie Review Dataset, is popular for binary sentiment classification. The notably brief reviews are highly polarized. While this is a large dataset, we only use 4503 reviews to mimic the size of the other two datasets that have depression screening labels.

B. Experimental Setup

We implement our nine cSeqGAN architectures within the Texygen benchmarking platform [42] for unconditioned text generation. Additionally, We also generate text with unconditioned SeqGAN models as baselines for comparison; a different SeqGAN model is required for each class. To train the generative models, we down sample each dataset to the size of the minority class: 2133 for movie reviews, 1207 interview replies, and 2680 for text messages. The discriminators are pretrained for 50 epochs before the 50 adversarial training epochs. Each model generate 4480 labeled texts. We run each model five times to obtain a confidence interval. The models were run on an internal cluster using NVIDIA V100 or T4 GPUs with at most 64 megabytes of RAM.

C. Machine Experimental Evaluation Approach

We evaluate the generated text quality with two established and popular text generation metrics [42]. The first of these metrics is the negative log-likelihood (NLL_{gen}) which is an output of the recurrent generator. A lower NLL_{gen} is indicative of better generated sentence diversity. In contrast, the Bilingual Evaluation Understudy (BLEU) score [50] assesses the similarity of the generated sentences with the real sentences. Effectively, a higher score is indicative of more realistic text. Given the short length of our input data, we focus our analysis on the BLEU-2 scores which assesses 2-gram matches, though we also report on BLEU-3 and BLEU-4 scores.

We further compare the predictive value of the real text and generated text. Bidirectional Encoder Representations from Transformers (BERT) [39] is a pretrained language representation model that can be fine-tuned for many tasks. The pretraining on a large corpus allows BERT to create useful feature embeddings for smaller datasets. Previously, BERT classifiers have proven effective at classifying short texts, such as sentiment from movie reviews [51], disaster events from tweets [52], and depression from clinical interview transcripts [4]. Thus, we use BERT classifiers with the parameters successful in related research [51]: learning rate of $2x10^{-5}$, 4 training epochs, and batch size of 32.

We only consider texts with at least two words for BERT input. We reserve 300 positive and 300 negative instances from each real dataset to use as testing data. As only four of the 135 conditional generative model runs failed to generate instances of each class with sufficient length, we proceed with 1200 (or the minimum class count) randomly sampled instances from each class as training data. For the real data and SeqGANs output, we also randomly sample 1200 instances of each class.

As we ran each conditional model five times, we have five NLL_{gen}, BLEU-2, BLEU-3, BLEU-4, and accuracy scores. The average and standard deviation of these scores are reported in Tables I-III. For the unconditioned SeqGAN models, we average the NLL_{gen} scores for all ten models (5 positive and 5 negative). To calculate BLEU-2 for SeqGAN, we combine the positive and negative output from two models.

D. Human Experimental Evaluation Approach

We further conduct Turing tests to evaluate the texts by having humans rate the text samples. While the related literature [24], [29] only uses Turing tests to assess text quality, we also have humans assess the text *predictiveness*. Thus, in addition to determine if text samples seemed realistic, evaluators were also tasked to determine what class they believed the sample belonged to; i.e., whether a sample of generated text originated from someone who screened positive for depression or not. We expect determining predictiveness to be difficult for the text with depression labels, due to ascertaining depression from only text message data is naturally expected to be a difficult task for human evaluators.

For each dataset, we construct five surveys consisting of 66 text samples. Only samples with more than two words were eligible for the surveys. In addition to the real data, we randomly sample three positive texts and three negative texts from the output of each of the ten generative models. As we ran each model five times, we thus form five surveys per dataset. Given the five surveys and six text samples per model, each model configuration was assessed by 30 text samples.

For each sample in the surveys, we ask two binary questions. The first assesses the predictive quality and the second assess the realism. For example, the first question for text messages asked *"Is the writer of this text message depressed or not depressed?"* with options "depressed" or "not depressed". The second question asked *"Was this message created by a human or a computer?"* with options "human" or "computer".

We recruited 32 university students to evaluate the texts under WPI IRB 00007374. Each of the 30 samples was rated by 3 students, resulting in 90 assessments for each model configuration. Text examples from these surveys are in Tables IV-VI. In each survey, we calculate the accuracy of responses for each model. The average of these five accuracies for each model configuration are reported in the aforementioned Tables.

V. Comparative Study Results

The results of our machine evaluation are in Tables I-III while the results of our human evaluation are in Tables IV-VI. Unlike for the conditioned models, we needed a separate unconditioned SeqGAN model for each class. While the other machine evaluation metrics were not applicable for the real text, we did calculate the accuracies of the real text for comparison.

A. Machine Evaluation: Text Quality and Diversity

From the results, we observe that the average BLEU-2 scores are higher for all of the cSeqGAN models than the unconditioned SeqGAN models for each dataset. The differences are largest for the clinical interviews in Table I where each of the conditional models have an average BLEU-2 score more than 0.3 higher than the unconditioned models. This indicates that the conditioned models produced more realistic text, likely due to parameter sharing.

When considering only the impact of weighting strategies and feedback mechanisms on the evaluation metrics, some patterns emerge. Notably, the lowest NLL_{gen} score was achieved by the models using the sentence weighting strategy. In contrast, the feedback mechanisms that yield the lowest NLL_{gen} scores for all weighting strategies are different for each dataset: single task for the movie reviews, dual task for the clinical interviews, and dual critic for text messages.

Comparing the BLEU scores of the nine cSeqGAN architectures also reveal some patterns. Dual unit weighting paired with dual task feedback has very high average BLEU-2 scores for clinical interviews and movie reviews while single unit weighting and single task feedback has very high average BLEU-2 scores for text messages and movie reviews. For BLEU-3 and BLEU-4, the single unit weighting strategy produced the most realistic text, especially when paired with the dual critic feedback mechanism.

B. Machine Evaluation: Text Predictiveness

Unfortunately, the generated text from the cSeqGAN models were not particularly predictive. For all three datasets, the

TABLE I: **Clinical Interviews:** Machine evaluation average ± standard deviation. Accuracy is obtained with BERT classifiers.

Weighting	Feedback	NLL	BLEU-2	BLEU-3	BLEU-4	Accuracy
Unconditioned	Unconditioned	**0.710** ± 0.084	0.203 ± 0.008	0.284 ± 0.008	0.164 ± 0.005	**0.548** ± 0.004
Sentence	Single Task	0.723 ± 0.013	0.544 ± 0.013	0.255 ± 0.020	0.143 ± 0.025	0.500 ± 0.008
Sentence	Dual Task	0.720 ± 0.020	0.550 ± 0.046	0.291 ± 0.032	0.163 ± 0.017	0.500 ± 0.008
Sentence	Dual Critic	0.708 ± 0.024	0.504 ± 0.029	0.271 ± 0.008	0.155 ± 0.005	0.504 ± 0.007
Single Unit	Single Task	0.733 ± 0.032	0.527 ± 0.040	0.262 ± 0.023	0.158 ± 0.015	0.512 ± 0.006
Single Unit	Dual Task	0.767 ± 0.032	0.516 ± 0.028	**0.298** ± 0.040	**0.172** ± 0.026	0.512 ± 0.009
Single Unit	Dual Critic	0.756 ± 0.018	0.539 ± 0.037	0.279 ± 0.027	0.163 ± 0.012	0.498 ± 0.004
Dual Unit	Single Task	0.740 ± 0.026	0.518 ± 0.023	0.282 ± 0.038	0.164 ± 0.024	0.506 ± 0.004
Dual Unit	Dual Task	0.768 ± 0.074	**0.556** ±0.035	0.270 ± 0.021	0.163 ± 0.015	0.504 ± 0.007
Dual Unit	Dual Critic	0.757 ± 0.030	0.504 ± 0.028	0.081 ± 0.002	0.051 ± 0.000	0.494 ± 0.006

TABLE II: **Text Messages:** Machine evaluation average ± standard deviation. Accuracy is obtained with BERT classifiers.

Weighting	Feedback	NLL	BLEU-2	BLEU-3	BLEU-4	Accuracy
Unconditioned	Unconditioned	0.247 ± 0.052	0.223 ± 0.060	0.155 ± 0.021	0.100 ± 0.012	**0.674** ± 0.018
Sentence	Single Task	0.216 ± 0.005	0.335 ± 0.035	0.153 ± 0.012	0.105 ± 0.009	0.487 ± 0.033
Sentence	Dual Task	**0.214** ± 0.012	0.315 ± 0.018	0.146 ± 0.010	0.096 ± 0.006	0.520 ± 0.029
Sentence	Dual Critic	0.220 ± 0.004	0.317 ± 0.032	0.153 ± 0.012	0.098 ± 0.009	0.478 ± 0.039
Single Unit	Single Task	0.229 ± 0.006	**0.340** ± 0.019	0.156 ± 0.013	0.102 ± 0.007	0.469 ± 0.041
Single Unit	Dual Task	0.225 ± 0.009	0.326 ± 0.014	0.134 ± 0.018	0.086 ± 0.015	0.515 ± 0.039
Single Unit	Dual Critic	0.234 ± 0.009	0.306 ± 0.034	**0.160** ± 0.005	**0.105** ± 0.005	0.506 ± 0.036
Dual Unit	Single Task	0.230 ± 0.008	0.325 ± 0.020	0.146 ± 0.013	0.100 ± 0.014	0.503 ± 0.036
Dual Unit	Dual Task	0.225 ± 0.003	0.305 ± 0.033	0.152 ± 0.008	0.101 ± 0.008	0.509 ± 0.023
Dual Unit	Dual Critic	0.235 ± 0.004	0.336 ± 0.018	0.083 ± 0.018	0.054 ± 0.011	0.506 ± 0.031

TABLE III: **Movie Reviews:** Machine evaluation average ± standard deviation. Accuracy is obtained with BERT classifiers.

Weighting	Feedback	NLL	BLEU-2	BLEU-3	BLEU-4	Accuracy
Unconditioned	Unconditioned	**1.736** ± 0.073	0.269 ± 0.021	0.155 ± 0.01	0.088 ± 0.004	**0.773** ± 0.016
Sentence	Single Task	2.072 ± 0.055	0.370 ± 0.017	0.157 ± 0.007	0.087 ± 0.003	0.490 ± 0.019
Sentence	Dual Task	2.055 ± 0.065	0.371 ± 0.025	0.156 ± 0.014	0.09 ± 0.007	0.519 ± 0.016
Sentence	Dual Critic	2.016 ± 0.071	0.384 ± 0.019	0.158 ± 0.016	0.09 ± 0.007	0.496 ± 0.042
Single Unit	Single Task	2.286 ± 0.120	0.395 ± 0.009	0.151 ± 0.009	0.083 ± 0.006	0.506 ± 0.032
Single Unit	Dual Task	2.185 ± 0.044	0.387 ± 0.008	0.163 ± 0.004	0.092 ± 0.003	0.489 ± 0.043
Single Unit	Dual Critic	2.229 ± 0.031	0.359 ± 0.033	**0.167** ± 0.009	**0.093** ± 0.004	0.516 ± 0.030
Dual Unit	Single Task	2.266 ± 0.024	0.380 ± 0.025	0.145 ± 0.014	0.083 ± 0.005	0.483 ± 0.032
Dual Unit	Dual Task	2.254 ± 0.053	**0.397** ± 0.011	0.157 ± 0.003	0.087 ± 0.003	0.492 ± 0.033
Dual Unit	Dual Critic	2.174 ± 0.075	0.374 ± 0.016	0.091 ± 0.013	0.056 ± 0.006	0.479 ± 0.019

generated text from the unconditioned models performed better in the BERT classifiers than the text from the conditioned models. Interestingly, the depression screening ability of the generated interview transcripts from all models exceeded that of the real interview transcripts which achieved an unexpectedly low accuracy of 0.485 ± 0.004; this indicates that the generative models amplified the signal for the class label. However, this is not true for the text messages and movie reviews where the real data achieved respectable accuracies of 0.711 ± 0.027 and 0.831 ± 0.015, respectively. Thus, the real data proved more predictive than the generated data for these two datasets.

C. Human Evaluation

As is the standard in unconditioned text generation research [24], [29], [27], [30], we tasked our human evaluators with assessing the realness of the generated texts. For each dataset, our evaluators understandably achieved the highest accuracy on the real data: 0.800 ± 0.056 for clinical interviews, 0.856 ± 0.044 for text messages, and 0.767 ± 0.108 for movie reviews. Further, the samples from the unconditioned models were not rated the most or least realistic. Notably, for the depression datasets, the text from the sentence weighting strategy paired with the single task feedback mechanism was rated least realistic.

We also tasked our human evaluators with assessing the predictive value of the generated texts, which we anticipated to be very difficult for depression detection. This hypothesis was validated, as the highest accuracies were 0.60 and 0.57 for the depression datasets. Unexpectedly, the accuracies on the real data was even lower with 0.567 ± 0.089 for clinical interviews and 0.522 ± 0.097 for text messages. Thus, for the real text messages, BERT classifiers were able to detect a depression signal that the human evaluators were not.

TABLE IV: **Clinical Interviews:** Accuracy average ± standard deviation for the human evaluation tasks and survey examples.

		Realness	Predictiveness	Not Depressed Example	Depressed Example
Unconditioned	Unconditioned	0.689 ± 0.143	0.578 ± 0.044	whats a valuable lately um	end i bad awake reading
Sentence	Single Task	0.500 ± 0.136	0.467 ± 0.232	and i dont know i am	um a person of different thing
Sentence	Dual Task	0.600 ± 0.206	0.567 ± 0.065	oh sniff that you know	or not much how so
Sentence	Dual Critic	**0.745** ± 0.129	0.589 ± 0.109	uh the first one	ive been feeling pretty um
Single Unit	Single Task	0.622 ± 0.231	0.544 ± 0.108	ten imaginable consultant part	uh that was about
Single Unit	Dual Task	0.622 ± 0.065	0.456 ± 0.082	hanging with friends so	so im sorry
Single Unit	Dual Critic	0.522 ± 0.083	0.444 ± 0.121	uh two ago more	stress in my mother
Dual Unit	Single Task	0.645 ± 0.156	0.444 ± 0.035	every few years ago	i dont felt members and groups
Dual Unit	Dual Task	0.622 ± 0.187	0.500 ± 0.117	just dont be more	uh i dont dont like
Dual Unit	Dual Critic	0.711 ± 0.181	**0.600** ± 0.102	im really really good for me	i think like down

TABLE V: **Text Messages:** Accuracy average ± standard deviation for the human evaluation tasks and survey examples.

		Realness	Predictiveness	Not Depressed Example	Depressed Example
Unconditioned	Unconditioned	0.500 ± 0.099	0.533 ± 0.125	im gon na ease up it	how like they want you too
Sentence	Single Task	0.278 ± 0.182	0.533 ± 0.156	i am home	it wasnt a bad table
Sentence	Dual Task	0.400 ± 0.223	0.433 ± 0.065	i was text u up	miss u too stumbling
Sentence	Dual Critic	0.511 ± 0.177	**0.567** ± 0.042	its going to seaside lol	sure call about 200 lol
Single Unit	Single Task	0.433 ± 0.178	0.500 ± 0.070	them gon na control	oh we was gon na play
Single Unit	Dual Task	0.456 ± 0.249	0.522 ± 0.147	ill ask some for dakota	your still waiting on it
Single Unit	Dual Critic	0.511 ± 0.154	0.533 ± 0.075	do im sure love u	u want to be there
Dual Unit	Single Task	0.444 ± 0.126	0.500 ± 0.070	much better but if im ok	last way is not ok
Dual Unit	Dual Task	**0.589** ± 0.178	0.511 ± 0.042	it went had a wonderful birthday	yeah babe just getting mad than
Dual Unit	Dual Critic	0.478 ± 0.120	0.500 ± 0.061	idk if ill be out then	it warm had is sad

TABLE VI: **Movie Reviews:** Accuracy average ± standard deviation for the human evaluation tasks and survey examples.

		Realness	Predictiveness	Positive Sentiment Example	Negative Sentiment Example
Unconditioned	Unconditioned	0.656 ± 0.089	**0.700** ± 0.188	a moving big drama with cool	we both scary
Sentence	Single Task	0.622 ± 0.022	0.411 ± 0.114	and theyre not an actor	eerily accurate depiction of admission
Sentence	Dual Task	0.500 ± 0.149	0.611 ± 0.208	just like how bad	great movies have no documentary
Sentence	Dual Critic	0.500 ± 0.099	0.511 ± 0.133	i hate a sound movie	its too bad to watch
Single Unit	Single Task	0.589 ± 0.075	0.322 ± 0.155	all in a terrific thing	but it feels strangely diverting
Single Unit	Dual Task	**0.689** ± 0.155	0.511 ± 0.074	its really dull	but it never is really funny
Single Unit	Dual Critic	0.611 ± 0.157	0.478 ± 0.171	the pool drowned me	dull or tuned and entertaining
Dual Unit	Single Task	0.467 ± 0.167	0.411 ± 0.264	you like the first enjoyable movie	a modernday point
Dual Unit	Dual Task	0.567 ± 0.124	0.378 ± 0.108	a pleasure of fiction	who are boring
Dual Unit	Dual Critic	0.533 ± 0.152	0.567 ± 0.226	pompous and good documentary	its not too immature and unpleasant

Our evaluators were better at classifying the intentionally polarizing movie reviews with an accuracy of 0.811 ± 0.097 on the real text, which is similar to the 0.831 ± 0.0015 achieved by the BERT classifiers. Of the generated movie reviews, the unconditioned models had the highest accuracy of 0.70.

VI. Discussion, Limitations, & Future Work

A. Contributions

In this research, we identified three weighting strategies and three feedback mechanisms for conditional adversarial text generation. These approaches combine to create nine unique cSeqGAN architectures. We further conduct a comprehensive comparative evaluation of these conditioning approaches on three small text datasets with depression and sentiment labels. As the first comparative study for fundamental conditional text generation strategies, we provide a valuable resource to inform future text generation applications and research. Our study is particularly useful for the healthcare domain where datasets tend to be small [5] and can therefore benefit from augmentation.

B. Data Limitations

The BERT classifiers were unfortunately unable to classify the real clinical interview sentences. Unlike prior work that successfully classifies DAIC-WOZ transcripts with BERT [4], [53], [54], we made the task more difficult by combining the responses to all questions in a single corpus and treating each sentence as a separate instance. Despite this preprocessing to increase the number of data instances, the clinical interviews remained the smallest dataset. Additionally, the clinical interviews were the least polarized dataset, as the movie reviews were intentionally polarized [45] and our preprocessing of the text messages involved only using texts sent by participants with the most polarizing depression screening scores. Unlike for the transcripts, the BERT classifiers were able to classify movie review and text message datasets which achieved accuracies of 0.831 and 0.711 respectively.

Additionally, the variety of the data we worked with exacerbated the issue of small dataset size. Prior work on short-text stylometry has shown that authors can be identified from even short messages, such as emails and texts [55]. This

indicates that individuals have unique styles when composing text messages and movie reviews. Thus, considering there were only a handful of samples from each user, the models struggled to generalize and accurately recreate their styles. This problem may be lessened if the vocabulary size was limited. Future work on the trade-off between vocabulary size and the dataset volume needed to train effective generative models is a promising direction for future work to build off of our results.

C. Performance Trade-off

It is unfortunate that the machine evaluation metrics indicate that there is a trade-off between generating realistic texts and generating predictive texts. While we selected a BERT classifier based on its use in prior depression detection research with small datasets [4], [16], it is possible that a different classifier would not result in this trade-off. This in in fact a research direction unto itself. Given our results, neither the unconditioned nor conditioned models currently perform sufficiently for their generated texts to be useful in augmenting the small existing real datasets for the purpose of depression screening. Yet, our comprehensive comparative study promises to help further research in this domain and can be built off of to yield a more viable solution in the future.

D. Future Opportunities

In this paper, we focused on generating text for depression detection. Thus, we demonstrated the conditional models on the small datasets available in this domain. Yet, our conditioning strategies are also applicable for larger labeled text datasets. While we implemented our nine cSeqGAN architectures to generate data with binary labels, all of the proposed conditional models are also easily scalable for more labels as there are no extra components that are required for extending the label set. Further our weighting strategies can be applied to any generative model with a recurrent network as a generator and our feedback mechanisms can be applied with any discriminator. Since we implemented the cSeqGAN architectures within the Texygen benchmarking platform [42], it would be easy to apply our conditioning approaches to the other generative networks.

VII. Conclusion

Due to small datasets being very common within healthcare [5], generating labeled text data for augmentation has the potential to greatly improve diagnostic and prognostic modeling. To this end, we conduct the first comparative study of conditional adversarial networks for generating small text datasets typical of the healthcare domain. In particular, we assembled a family of nine cSeqGAN models with unified architectures that make them applicable for smaller datasets and scalable regardless of the number of classes. We then use our cSeqGAN models to generate labeled transcripts, text messages, and movie reviews. In addition to determining quality, we also use both machine and human assessments to determine the usefulness of the generated text to detect depression and sentiment. As we implemented all of our cSeqGAN models within a unified text generation benchmarking platform, they are a valuable resource for both machine learning and healthcare researchers.

Acknowledgment

We thank the U.S. Department of Education P200A150306 & P200A180088: GAANN grants, NSF REU site 1560229, NSF III: Small #1910880, NSF III: Small #1815866, and AFRI Grant 1023720 for the funding that made this work possible. Results were obtained in part using a high-performance computing cluster acquired through NSF MRI DMS-1337943 to WPI.

We thank Mahum Shah, Nicholas Pingal, Karthika Suresh, Matthew Dzwil, Luke Buquicchio, Ermal Toto, other Emutivo researchers, and the DAISY lab at WPI for their advice, feedback and support during the development of this work. This research was completed in part while ML Tlachac was a PhD candidate at WPI and Benjmain Litterer and Saitheeraj Thatigotla were at WPI for NSF REU.

References

[1] World Health Organization *et al.*, "The who special initiative for mental health (2019-2023): universal health coverage for mental health," World Health Organization, Tech. Rep., 2019.

[2] J. Torous, M. V. Kiang, J. Lorme, J.-P. Onnela *et al.*, "New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research," *JMIR mental health*, vol. 3, no. 2, p. e5165, 2016.

[3] M. L. Tlachac and E. Rundensteiner, "Screening for depression with retrospectively harvested private versus public text," *IEEE journal of biomedical and health informatics*, vol. 24, no. 11, pp. 3326–3332, 2020.

[4] E. Toto, M. L. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, 2021, pp. 4145–4154.

[5] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, 2018.

[6] J. Guan, R. Li, S. Yu, and X. Zhang, "Generation of synthetic electronic medical record text," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 374–380.

[7] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing (UbiComp)*, 2014, pp. 3–14.

[8] M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes, "Demonicsalmon: Monitoring mental health and social interactions of college students using smartphones," *Smart Health*, vol. 9, pp. 192–203, 2018.

[9] A. Dogrucu, A. Perucic, A. Isaro, D. Ball, E. Toto, E. A. Rundensteiner, E. Agu, R. Davis-Martin, and E. Boudreaux, "Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data," *Smart Health*, vol. 17, p. 100118, 2020.

[10] S. Ware, C. Yue, R. Morillo, J. Lu, C. Shang, J. Kamath, A. Bamis, J. Bi, A. Russell, and B. Wang, "Large-scale automatic depression screening using meta-data from wifi infrastructure," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 2, no. 4, pp. 1–27, 2018.

[11] D. Di Matteo, K. Fotinos, S. Lokuge, J. Yu, T. Sternat, M. A. Katzman, J. Rose *et al.*, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study," *JMIR Form. Res.*, vol. 4, no. 8, 2020.

[12] M. L. Tlachac, E. Toto, J. Lovering, R. Kayastha, N. Taurich, and E. Rundensteiner, "Emu: Early mental health uncovering framework and dataset," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) Special Session Machine Learning in Health*. IEEE, 2021, pp. 1311–1318.

[13] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

[14] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu *et al.*, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 1, pp. 1–41, 2021.

[15] M. L. Tlachac, R. Flores, M. Reisch, K. Houskeeper, and E. Rundensteiner, "DepreST-CAT: Retrospective smartphone call and text logs collected during the covid-19 pandemic to screen for mental illnesses," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 2, pp. 1–32, 2022.

[16] M. L. Tlachac, R. Flores, M. Reisch, R. Kayastha, N. Taurich, V. Melican *et al.*, "StudentSADD: Rapid mobile depression and suicidal ideation screening of college students during the coronavirus pandemic," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 2, pp. 1–32, 2022.

[17] W. Wang, S. Nepal, J. F. Huckins, L. Hernandez, V. Vojdanovski, D. Mack *et al.*, "First-gen lens: Assessing mental health of first-generation students across their first year at college using mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 2, pp. 1–32, 2022.

[18] C. Andrade, "Sample size and its importance in research," *Indian journal of psychological medicine*, vol. 42, no. 1, pp. 102–103, 2020.

[19] M. Abedi, L. Hempel, S. Sadeghi, and T. Kirsten, "Gan-based approaches for generating structured data in the medical domain," *Applied Sciences*, vol. 12, no. 14, p. 7075, 2022.

[20] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[21] J. Yoon, L. N. Drumright, and M. Van Der Schaar, "Anonymization through data synthesis using generative adversarial networks (ads-gan)," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2378–2388, 2020.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[24] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[25] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, and Y. Bengio, "Maximum-likelihood augmented discrete generative adversarial networks," *arXiv preprint arXiv:1702.07983*, 2017.

[26] M. J. Kusner and J. M. Hernández-Lobato, "Gans for sequences of discrete elements with the gumbel-softmax distribution," *arXiv preprint arXiv:1611.04051*, 2016.

[27] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," *arXiv preprint arXiv:1705.11001*, 2017.

[28] X. Zhang and Y. LeCun, "Text understanding from scratch," *arXiv preprint arXiv:1502.01710*, 2015.

[29] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[30] W. Nie, N. Narodytska, and A. Patel, "Relgan: Relational generative adversarial networks for text generation," in *International conference on learning representations (ICLR)*, 2018.

[31] K. Wang and X. Wan, "Sentigan: Generating sentimental texts via mixture adversarial networks," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 4446–4452.

[32] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 1587–1596.

[33] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A generative model for category text generation," *Information Sciences*, vol. 450, pp. 301–315, 2018.

[34] Z. Liu, J. Wang, and Z. Liang, "Catgan: Category-aware generative adversarial networks with hierarchical evolutionary learning for category text generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8425–8432.

[35] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable gans," *arXiv preprint arXiv:2101.02477*, 2021.

[36] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 68–81, 2019.

[37] R. Kitchin and G. McArdle, "What makes big data, big data? exploring the ontological characteristics of 26 datasets," *Big Data & Society*, vol. 3, no. 1, p. 2053951716631130, 2016.

[38] M. Tlachac, V. Melican, M. Reisch, and E. Rundensteiner, "Mobile depression screening with time series of text logs and call logs," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 4171–4186.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[41] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, and L. Carin, "Adversarial feature matching for text generation," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 4006–4015.

[42] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, "Texygen: A benchmarking platform for text generation models," in *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1097–1100.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[45] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *49th Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 142–150.

[46] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014.

[47] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[48] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[49] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[51] S. A. Rauf, Y. Qiang, S. B. Ali, and W. Ahmad, "Using bert for checking the polarity of movie reviews," *International Journal of Computer Applications*, vol. 975, 2019.

[52] H. M. Zahera, I. A. Elgendy, R. Jalota, M. A. Sherif, E. Voorhees, and A. Ellis, "Fine-tuned bert model for multi-label tweets classification," in *Proceedings of the 28th Text REtrieval Conference (TREC)*, 2019, pp. 1–7.

[53] R. Flores, M. L. Tlachac, E. Toto, and E. Rundensteiner, "Transfer learning for depression screening from follow-up clinical interview questions," *Deep Learning Applications*, vol. 4, 2022, in Press.

[54] S. Senn, M. L. Tlachac, R. Flores, and E. Rundensteiner, "Ensembles of bert for depression classification," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 4691–4694.

[55] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," in *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2013, pp. 1–6.