# GAN-BERT for Automated Essay Scoring

Griffin Holt and Theodore Kanell
gholt@stanford.edu, tkanell@stanford.edu

## Problem

Every year, millions of individuals take English language proficiency exams, such as TOEFL and IELTS, for professional and academic development. These exams are typically graded by human evaluators; automating the evaluation process can improve both efficiency and fairness of the examinations. Our approach to the Automated Essay Scoring (AES) task is to implement three variations of the GAN-BERT architecture: a feed-forward neural network generator; a BERT transformer generator; and a generator composed of a fine-tined GPT2 language model in tandem with a BERT transformer. We use a single pre-trained RoBERTa model, fine-tuned to our task and dataset, for a baseline comparison.

## Background

Formally, the Automated Essay Scoring (AES) task is defined as follows: given an essay with $m$ words $X = \{x_i\}_{i=1}^{m}$, we want to output a single score $y$ that reflects the measure of the essay.

Previous attempts [3] to create an effective and accurate AES system followed two basic designs: deep neural network models using either LSTM or CNN architectures using factors such as word length, spelling errors, or bag of words to featurize essays in a time consuming procedure; and transformer-based models, such as BERT.

Very few researchers have applied GAN networks to NLP tasks and none have applied it to the AES task. Croce et al. [2] was able to achieve state-of-the-art results in Question Classification, Community Question-Answering, and Argument Boundary detection. Croce et al. [2] also demonstrated that applying a semi-supervised GAN on a NLP task can enable the model to achieve high results with far fewer labeled data points on Sentiment Classification.

## Data

We utilized the ETS Corpus of Non-Native Written English [1], a compilation of 12,100 English essays written by speakers of 11 non-English native languages (1,100 essays for each language) across 8 different essay prompts as part of the international academic English language proficiency exam, TOEFL. The dataset was developed specifically for native language identification, but, as acknowledged by its authors, can be used for other tasks (such as AES).

Each essay $X_k$ is labeled with a human-evaluated score $y_k \in \{\text{Low}, \text{Medium}, \text{High}\} = \mathcal{S}$. The training set is composed of $n = 9900$ essays, and the development and test sets are each composed of $\tilde{n} = 1100$ essays.

## GAN-BERT Architecture

In our project, we extend the GAN-BERT architecture–a unique adaptation of the Generative Adversarial Network (GAN) that incorporates a BERT transformer and was first introduced by [2] for various NLP tasks–to the Automated Essay Scoring task.

Let $G$ denote the generator network and $D$ denote the discriminator network. Let $X_k = \{x_i\}_{i=1}^{m}$ represent a real essay from our dataset. Each essay $X_k$ is labeled with a human-evaluated score $y_k \in \{\text{Low}, \text{Medium}, \text{High}\} = \mathcal{S}$.

An essay $X_k$ is passed into a pre-trained BERT module which we fine-tuned in advance on the AES task. The BERT output $v_B = h_{CLS}$ is then passed into the discriminator $D$, a feed-forward neural network (see Figure 2d). The final output of the discriminator $D$ is the predicted class $\hat{y}_k$.

Separately, "noisy input" is passed into the generator $G$. For the AES task, we experiment with three different generator structures (see Figure 2). The exact definition of "noisy input" depends on the structure of the generator itself. Regardless of its internal structure, the generator $G$ produces an output $v_G$: a "fake" sample that, ideally, mimics the output $v_B$ of the BERT module when fed a real essay $X$. This generator output $v_G$ is then passed into the discriminator $D$ and assigned a class probability and final prediction $\hat{y}_{\tilde{k}}$.

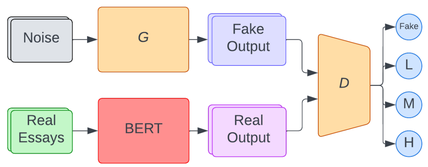
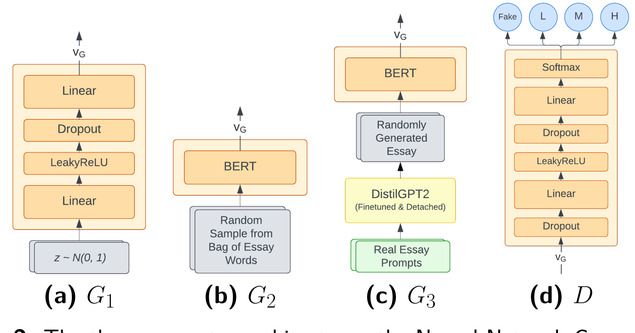
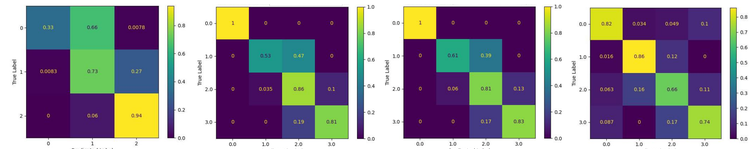



**Figure 1:** The GAN-BERT architecture, as described by [2]: essays are passed into the BERT transformer module, and the discriminator $D$ outputs the essay score $\hat{y} \in \{\text{Fake}, \text{L}, \text{M}, \text{H}\}$, where $\hat{y} = \text{Fake}$ signifies $D$ identified the input as generated the generator $G$.




**(a)** $G_1$ **(b)** $G_2$ **(c)** $G_3$ **(d)** $D$

**Figure 2:** The three generator architectures–the Neural Network Generator $G_1$; the BERT Generator $G_2$; and the GPT2-BERT Generator $G_3$–and the Discriminator $D$ architecture




**(a)** RoBERTa **(b)** $G_1$ **(c)** $G_2$ **(d)** $G_3$

**Figure 3:** Confusion Matrices for RoBERTa Baseline and GAN-BERT Model Performance on the Test Sets, normalized over the true counts (i.e., by row)

## Results

All three GAN-BERT architectures ($G_1$, $G_2$, $G_3$) acheived better performance for essay scoring than the baseline RoBERTa model. More specifically, all three GAN-BERT architectures improved upon the baseline RoBERTa model in being able to better differentiate between Low and Medium level essays, and between Medium and High level essays. The feed-forward generator $G_1$ and $G_2$ GAN-BERT models had the highest evaluation scores. The $G_3$ model demonstrated the most evidence of taking advantage of the competitive nature of the GAN structure: the generator confused the discriminator more often (one times out of ten) without compromising too much on scoring performance.

## Future Work

With additional training time, we would attach the DistilGPT2 module in the $G_3$ generator to the gradient (as opposed to fine-tuning it in advance and freezing its parameters during the GAN-BERT training).

We also suggest a fourth GAN-BERT structure $G_4$: the discriminator $D$ itself is a BERT transformer; essays are tokenized and fed directly into the discriminator; and the generator $G$ is a DistilGPT2 language model attached to the gradient.

Because of the unique expertise of each of our four types of models, we believe that a stacked ensemble or soft-voting ensemble of the four models (RoBERTA, $G_1$, $G_2$, and $G_3$) could achieve exceptionally high QWK scores and would be worth investigation.

## References

[1] Blanchard, Daniel, Tetreault, Joel, Higgins, Derrick, Cahill, Aoife, and Chodorow, Martin. 2014. Ets corpus of non-native written english ldc2014t06.

[2] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

[3] Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, Brussels, Belgium. Association for Computational Linguistics.