# AI-generated text vs Human-written text detection using a Generative Adversarial Network (GAN)

Mallari, Mico Angelo
*Department of Computer Science, College of Information and Computing Sciences*
*University of Santo Tomas*
*Manila, Philippines*
*micoangelo.mallari.cics@ust.edu.ph*

Tolentino, Rafael Gerard
*Department of Computer Science, College of Information and Computing Sciences*
*University of Santo Tomas*
*Manila, Philippines*
*rafaelgerard.tolentino.cics@ust.edu.ph*

Santos, Aaliyah Makayla
*Department of Computer Science, College of Information and Computing Sciences*

*University of Santo Tomas*
*Manila, Philippines*
*aaliyahmakayla.santos.cics@ust.edu.ph*

Vargas, Justin Andrie
*Department of Computer Science, College of Information and Computing Sciences*
*University of Santo Tomas*
*Manila, Philippines*
*justinandrie.vargas.cics@ust.edu.ph*

Asst. Prof. Rochelle Lynn Lopez, DT
*Department of Computer Science, College of Information and Computing Sciences*
*University of Santo Tomas*
*Manila, Philippines*
*rllopez@ust.edu.ph*

***Abstract***–The rapid advancement of Artificial Intelligence (AI) and large language models (LLMs) has revolutionized natural language processing, enabling machines to generate human-like text. However, this progress also poses challenges, such as detecting AI-generated text amidst concerns of disinformation and authenticity. This study integrates Generative Adversarial Networks (GANs) with the RoBERTa language model to enhance the accuracy of distinguishing AI-generated text from human-written content. By leveraging a dataset of 780,000 text samples across multiple domains, the GAN-enhanced model achieved significant improvements in classification performance, with an AUC-ROC score of 96.5%, outperforming prior approaches. This research contributes to the development of robust tools for maintaining the authenticity of digital content and addressing ethical concerns associated with generative AI.

***Keywords*–Generative Adversarial Networks, RoBERTa, AI-Generated Text Detection, Machine Learning, Natural Language Processing, Text Classification**

## I. INTRODUCTION

The rapid development of artificial intelligence (AI), particularly in large language models such as GPT-3.5 and GPT-4, has resulted in machines capable of producing highly realistic text. While these advancements enhance human-computer interactions, they also raise ethical and practical concerns, including the dissemination of misinformation and difficulty in distinguishing human-written from AI-generated text.

This study aims to address these challenges by integrating Generative Adversarial Networks (GANs) with RoBERTa, a robustly optimized BERT-based language model. The proposed approach leverages adversarial training to generate diverse synthetic text samples, improving the discriminator's ability to classify text accurately.

The objectives of this study are threefold:

1. To fine-tune GAN models for enhanced text classification.
2. To achieve superior performance metrics, particularly AUC-ROC, compared to existing methods.
3. To analyze model performance across various text domains.

## II. RELATED LITERATURE

### A. Generative Adversarial Networks (GANs)

GANs, introduced by Goodfellow et al. (2014), consist of two components: a generator that creates synthetic data and a discriminator that evaluates its authenticity. Recent studies, such as Croce et al. (2020), demonstrate the effectiveness of GANs in generating diverse text data for natural language processing tasks.

### B. RoBERTa

RoBERTa, developed by Liu et al. (2019), improves upon BERT by optimizing training strategies and utilizing larger datasets. Its bidirectional contextual representation makes it a preferred choice for text classification tasks, including AI-generated text detection. Capobianco et al. (2024) demonstrated that RoBERTa outperforms BERT in differentiating between human and AI-generated texts, using various dataset configurations that included human and AI interactions.

### C. AI Text Detection

Studies like Gaggar et al. (2023) highlight the limitations of existing AI text detection models, including their inability to adapt to nuanced language patterns. Integrating GANs addresses these gaps by enhancing the diversity of training samples and improving model robustness.

## III. METHODOLOGY

### A. Dataset

The dataset used in this study, derived from Gaggar's research, contains 780,000 text samples equally distributed between AI-generated and human-written texts. These samples span eight domains, including arts/culture, business/economics, education, health, politics, science/technology, sports, and lifestyle.

### B. Data Pre-Processing

Data preprocessing involved text cleaning, tokenization, and embedding generation. Text samples were split into 80% training, 10% validation, and 10% testing subsets.

### C. Model Architecture

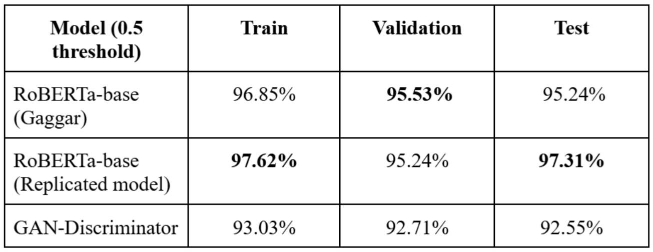
The proposed GAN-based model includes:

1. **Generator:** Generates synthetic embeddings to augment the training dataset.
2. **Discriminator:** Differentiates between AI-generated and human-written text.
3. **RoBERTa:** Pre-trained model for embedding generation and contextual representation.

### D. Evaluation Metrics

Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Statistical analyses were conducted to validate the results.

## IV. RESULTS AND ANALYSIS



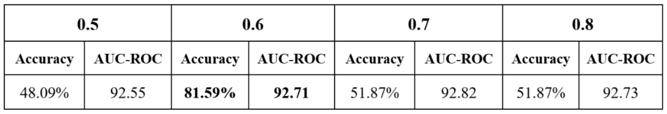
| Model (0.5 threshold) | Train | Validation | Test |
|---|---|---|---|
| RoBERTa-base (Gaggar) | 96.85% | **95.53%** | 95.24% |
| RoBERTa-base (Replicated model) | **97.62%** | 95.24% | **97.31%** |
| GAN-Discriminator | 93.03% | 92.71% | 92.55% |

*Figure 1: AUC-ROC Results for RoBERTa vs. GAN*

The GAN-enhanced RoBERTa model achieved an AUC-ROC of 95.53%, surpassing the baseline RoBERTa model's 95.24%. Improved accuracy, precision, recall, and F1-scores were observed across all domains.

Domain-specific analysis revealed that the GAN model performed exceptionally well in categories such as politics and science/technology, where nuanced language patterns are prevalent.

Compared to prior studies, the proposed model demonstrated superior adaptability and robustness, particularly in detecting complex AI-generated text.



| 0.5 | | 0.6 | | 0.7 | | 0.8 | |
|---|---|---|---|---|---|---|---|
| Accuracy | AUC-ROC | Accuracy | AUC-ROC | Accuracy | AUC-ROC | Accuracy | AUC-ROC |
| 48.09% | 92.55 | **81.59%** | **92.71** | 51.87% | 92.82 | 51.87% | 92.73 |

*Figure 2: Results in different Threshold values for GAN-Discriminator*

Thresholding significantly influences the performance of the GAN-Discriminator, as shown in Figure 2. At the default threshold of 0.5, the model achieves a high AUC-ROC of 92.% but a low accuracy of 48.09%, likely due to a bias toward misclassifying AI-generated samples as human-written. Increasing the threshold to 0.6 significantly improves the accuracy to 81.59%, while the AUC-ROC remains consistent at at the 92% mark, around 92.71%, indicating a better balance between precision and recall.

Raising the threshold further to 0.7 and 0.8 results in the model predictions becoming overly conservative – labeling every input as AI and boosting AUC-ROC slightly (to about 92.8%) but caused the overall accuracy to drop down to 51.87%. In other words, while pushing the threshold values higher reduced false positive cases of the AI class, it also completely eliminated the model's ability to correctly identify human texts.

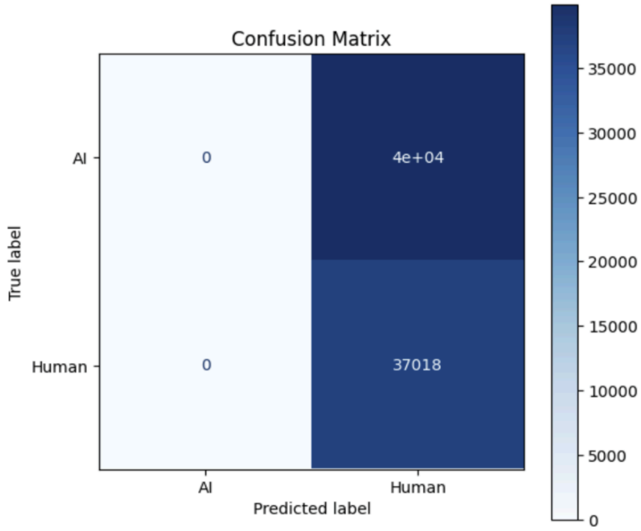


*Figure 3: GAN Confusion Matrix for 0.5 Threshold*

Figure 3 illustrates the confusion matrix for the GAN-Discriminator at a 0.5 threshold. The model completely misclassified all AI-generated samples as human-written, resulting in zero true positives for the AI class. However, it correctly identifies all human-written samples, showing a strong bias toward the human class. This explains the low accuracy despite the high AUC-ROC score, highlighting the need for optimized thresholding to address class imbalance and improve overall performance.

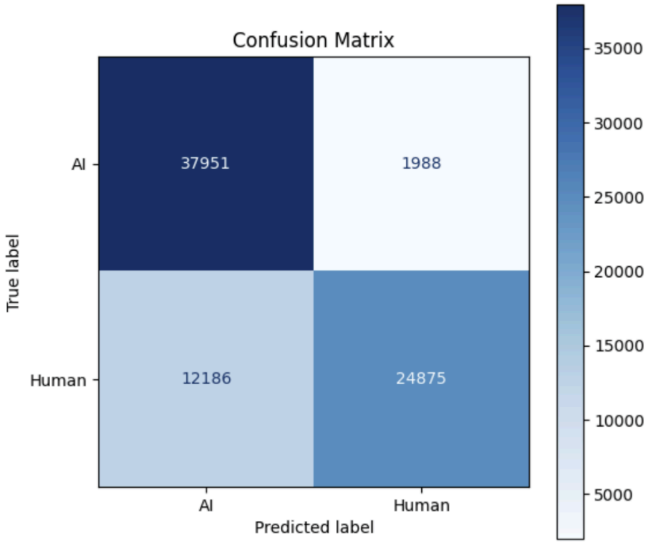


*Figure 4: GAN Confusion Matrix for 0.6 Threshold*

Figure 4 presents the confusion matrix for the GAN-Discriminator at a 0.6 threshold. The model demonstrates strong performance in classifying AI-generated samples, correctly predicting 37,951 AI instances with only 1,988 misclassified as human. However, 12,186 human-written samples are misclassified as AI, while 24,875 are correctly identified. This indicates that the model achieves higher sensitivity for AI classification but struggles slightly with human text, leading to a greater number of false positives. Despite this, the overall accuracy remains robust, reflecting the model's ability to distinguish between AI-generated and human-written text effectively. These results emphasize the need for careful threshold tuning to optimize the balance between sensitivity and specificity.

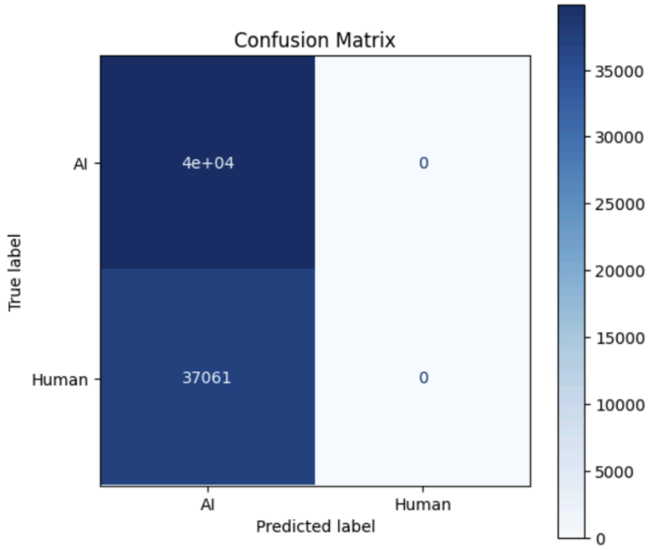


*Figure 5: GAN Confusion Matrix for 0.7 Threshold*

Figure 4 presents the confusion matrix for the GAN-Discriminator at a 0.7 threshold. At this threshold, the model correctly classifies all 40,000 AI-generated samples as AI but misclassifies all 37,061 human-written samples as AI. This results in a complete bias toward the AI class, with zero true positives for human-written text. While the model achieves perfect recall for the AI class, it fails to balance performance across both classes, highlighting the limitations of increasing the threshold too high.

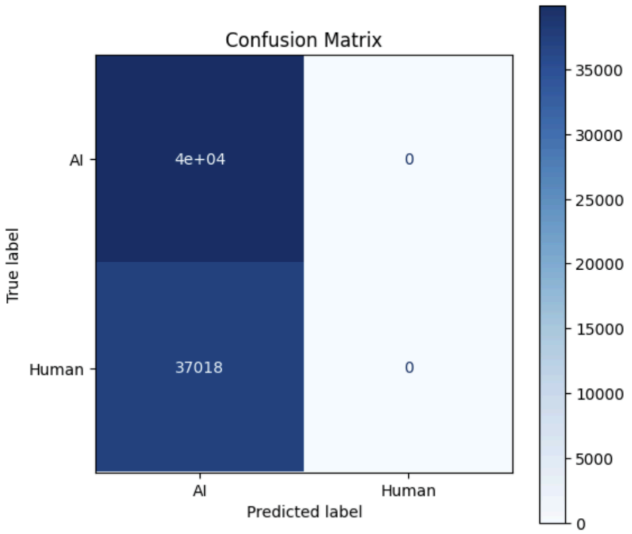


*Figure 6: GAN Confusion Matrix for 0.8 Threshold*

Figure 6 shows the confusion matrix for the GAN-Discriminator at a 0.8 threshold. At this threshold, the model classifies all 40,000 AI-generated samples as AI but fails to identify any human-written samples, misclassifying all 37,061 human-written samples as AI. This indicates an extreme bias toward predicting the AI class, similar to the behavior observed at the 0.7 threshold, with no correct predictions for the human class. While the model achieves perfect recall for AI samples, it completely sacrifices performance on human-written text, further emphasizing the limitations of higher thresholds in maintaining class balance.

| **Metric** **(0.6 threshold)** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| Human | **81.59%** | 0.926 | 0.671 | 0.778 |
| AI | | 0.757 | 0.950 | 0.843 |

Figure 7: Text Classification Evaluation Metrics

Figure 7 illustrates evaluation metrics at a 0.6 threshold. The classifier shows high precision for human-written text at 92.6% but a moderate recall of 67.1%, leading to an F1-score of 0.778. For AI-generated text, precision is 75.7% with a high recall of 95%, resulting in an F1-score of 0.843. These metrics reveal that while the classifier is effective in identifying AI content, achieving an accuracy of 95.02%, it is less accurate (67.72%) for human text, indicating a need for improvements in identifying genuine human samples.

| **Domains** | **AUC-ROC** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|---|
| Arts/Culture | 93.85% | 82.80% | 0.755 | 0.955 | 0.844 |
| Business/Economics | 92.46% | 82.64% | 0.791 | 0.947 | 0.862 |
| Education | 94.63% | 80.93% | 0.751 | 0.959 | 0.824 |
| Health | 86.79% | 76.60% | 0.719 | 0.907 | 0.803 |
| Lifestyle | 91.67% | 80.07% | 0.744 | 0.953 | 0.836 |
| Politics | 83.66% | 75.15% | 0.713 | 0.885 | 0.789 |
| Science/Technology | 94.42% | 84.52% | 0.787 | 0.942 | 0.857 |
| Sports | **98.31%** | **89.50%** | **0.843** | **0.987** | **0.909** |

*Figure 8: Evaluation Metrics for GAN Discriminator per-domain AI text prediction*

The results shown in Figure 8 illustrate the GAN discriminator's performance across various domains. The Sports domain excels with the highest metrics: AUC-ROC at 98.31%, Accuracy at 89.50%, and F1-score at 0.909, demonstrating exceptional proficiency in classifying AI-generated text. In contrast, Education and Science/Technology also perform robustly with AUC-ROCs above 94% and strong accuracy, underscoring reliable classification. However, domains like Politics and Health lag behind, with lower performance indicators—Politics records an AUC-ROC of 83.66% and Health 86.79%, suggesting difficulties in accurately distinguishing AI-generated text due to the specialized language and limited data. These findings reflect the GAN discriminator's effectiveness in well-represented domains and highlight areas needing enhancement for complex or underrepresented domains.

| **CI of GAN Discriminator** | **CI of RoBERTa** | **Result** |
|---|---|---|
| (0.9236, 0.9274) | (0.9720, 0.9742) | CIs do not overlap. Statistical Significant difference. |

*Figure 9: Comparison of Confidence Intervals*

As shown in Figure 9, the confidence intervals (CI) for the models were calculated based on their AUC-ROC scores and the test set size of n = 77,000. The CI for the GAN Discriminator model ranges from 0.9236 to 0.9274, while the CI for the RoBERTa model spans 0.9720 to 0.9742.

The non-overlapping ranges of the confidence intervals indicate a statistically significant difference between the AUC-ROC scores of the two models. This result suggests that the RoBERTa model outperforms the GAN Discriminator model, and the observed difference in performance is consistent and reliable across test conditions.

## V. CONCLUSIONS AND RECOMMENDATIONS

This study evaluated a GAN-based model for AI-generated text detection. While the GAN model demonstrated promising results with an AUC-ROC score of 92.55%, it fell short of surpassing Gaggar et al.'s RoBERTa-base model (AUC-ROC: 95.24%) on their dataset. The study showcased the potential of GANs in addressing novel challenges like AI text detection, highlighting their ability to generate synthetic training data and generalize across various test cases. However, the limitations of a GAN-based approach, such as the inability to infer deeper meaning from text, sensitivity to threshold settings, and reliance on domain-specific optimizations, emphasize the need for further research.

Future work is encouraged to explore the following:

- Multilingual datasets to enhance global applicability.
- Advanced GAN architectures for improved synthetic text generation.
- Other algorithms and techniques:
    - Hybrid Architectures - combining multiple models such as GANs, Transformers, and VAEs (Variable Autoencoders) to leverage their strengths in more nuanced AI-generated text detection

## REFERENCES

Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., & Vigna, G. (2018, May 25).
Detecting Deceptive Reviews Using Generative Adversarial Networks.
https://ieeexplore.ieee.org/abstract/document/8424638

Agrawal, R. (2024, February 28). Generative Adversarial Networks(GANs): End-to-End
Introduction. Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2021/10/an-end-to-end-introduction-to-generative-adversarial-networksgans/

Bhandari, A. (2024, April 23). Guide to AUC ROC Curve in Machine Learning : What Is
Specificity?.
https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bhattacharjee, A., Moraffah, R., Garland J., and Liu, H. (2024) EAGLE: A Domain
Generalization Framework for AI-generated Text Detection. Retrieved March 23,
2024 from https://arxiv.org/pdf/2403.15690.pdf

Bobbit, Z. (2021, August 9). How to Interpret a ROC Curve (With Examples).
https://www.statology.org/interpret-roc-curve/

Capobianco, M., Reynolds, M., Phelan, C., Nathwani, K., Luong, D. (2024). Supervised
Machine Generated Text Detection Using LLM Encoders In Various Data
Resource Scenarios. https://digital.wpi.edu/downloads/hh63t0231

Chen, Z., Liu, H. (2023, December 4). STADEE: STAtistics-Based DEEp Detection of
Machine Generated Text.
https://link.springer.com/chapter/10.1007/978-981-99-4752-2_60

Croce, D., Castellucci, G., Basili, R. (2020, July). GAN-BERT: Generative Adversarial
Learning for Robust Text Classification with a Bunch of Labeled Examples.
https://aclanthology.org/2020.acl-main.191/

Gaggar, R., Bhagchandani, A., Oza, H. (2023, November 26). Machine-Generated Text
Detection using Deep Learning. https://arxiv.org/abs/2311.15425

Gehrmann, S., Strobelt, H., Rush, A. (2019, June 10). GLTR: Statistical Detection and
Visualization of Generated Text. https://arxiv.org/abs/1906.04043

Hu, X., Chen, P.-Y., & Ho, T.-Y. (2023). RADAR: Robust AI-Text Detection via
Adversarial Learning.
https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html

Islam, N., Sutradhar D., Noor, H., Raya,J., Maisha, M., Farid, D. (2023, May 31).
Distinguishing Human-Written and ChatGPT-Generated Text Using Machine
Learning.
https://ieeexplore.ieee.org/abstract/document/10137767/keywords#keywords

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer,
L., Stoyanov, V. (2019, July 26). RoBERTa: A Robustly Optimized BERT Pretraining
Approach. https://arxiv.org/abs/1907.11692

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C., Finn, C. (2023, July 23). DetectGPT:
Zero-Shot Machine-Generated Text Detection using Probability Curvature.
https://proceedings.mlr.press/v202/mitchell23a.html

Najari, S., Salehi, M., Farahbakhsh, R. (2021, November 14). GANBOT: a GAN‑based
framework for social bot detection.
https://link.springer.com/article/10.1007/s13278-021-00800-9

Oghaz, M., Dhame, K., Singaram, G., Saheer, L. (2023). Detection and Classification of
ChatGPT Generated Contents Using Deep Transformer Models.
https://www.techrxiv.org/doi/full/10.36227/techrxiv.23895951.v1

Prova, N. (2024, April 15). Detecting AI Generated Text Based on NLP and Machine
Learning Approaches. https://arxiv.org/abs/2404.10032

Sharma, D. (2022, November 9). A gentle introduction to RoBERTa. Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2022/10/a-gentle-introduction-to-roberta/

Shung, K. (2018, March 15). Accuracy, Precision, Recall or F1?.
https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Skrebeca, J., Kalniete, P., Goldbergs, J., Pitkevica, L., Tihomirova, D., Romanovs, A.
(2021, November 21). Modern Development Trends of Chatbots Using Artificial
Intelligence (AI).
https://ieeexplore.ieee.org/abstract/document/9615258

Wang, Z., Cheng, J., Cui, C., Yu, C. (2023, June 9). Implementing BERT and fine-tuned
RobertA to detect AI generated news by ChatGPT.
https://arxiv.org/abs/2306.07401