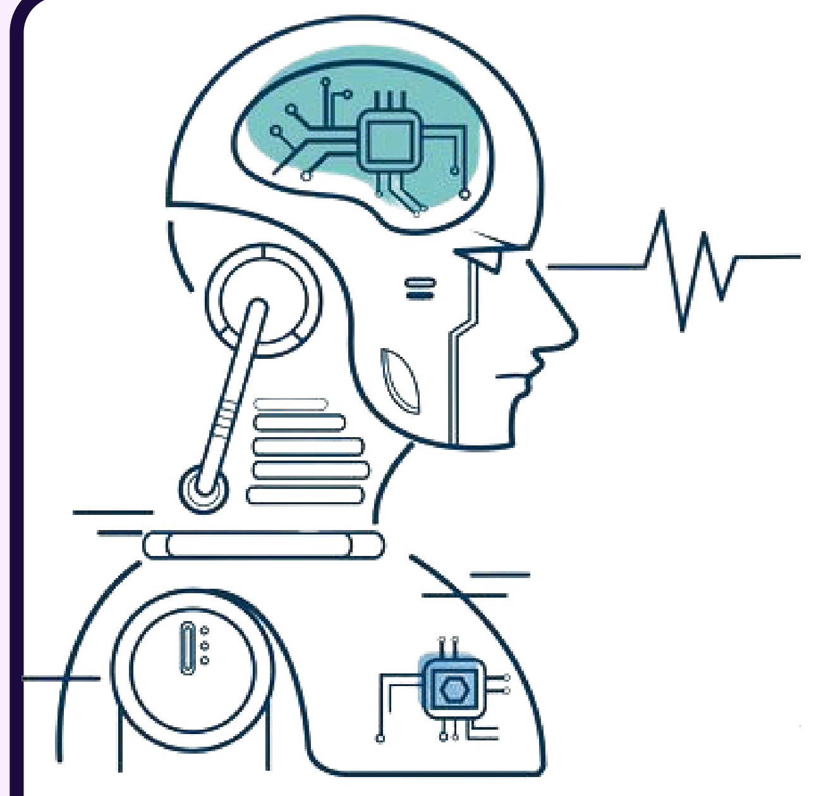
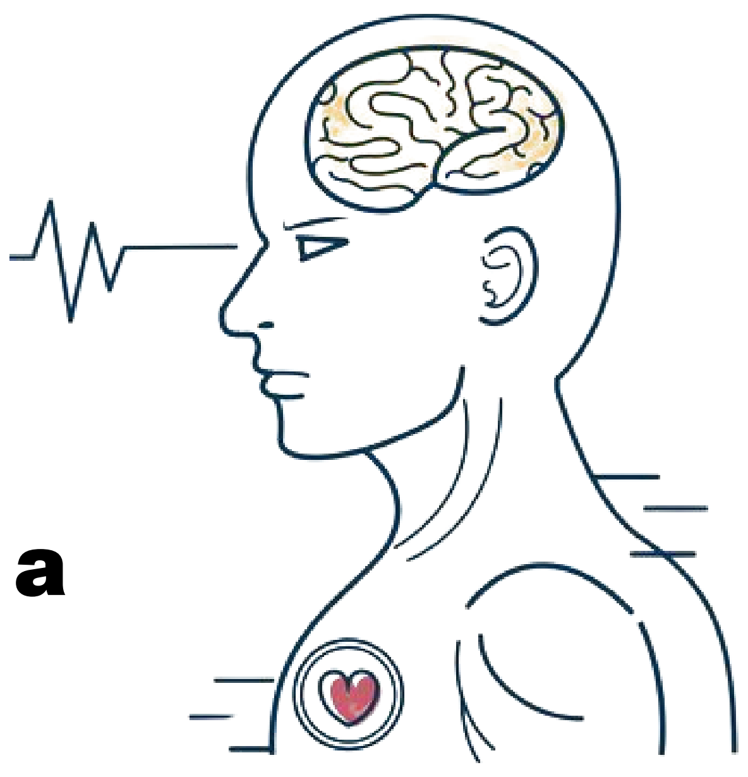


University of Santo Tomas
College of Information and Computing Sciences
Department of Computer Science



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

Mallari, Mico Angelo **Santos,** Aaliyah Makayla **Tolentino,** Rafael Gerard **Vargas,** Justin Andrie

A. 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.

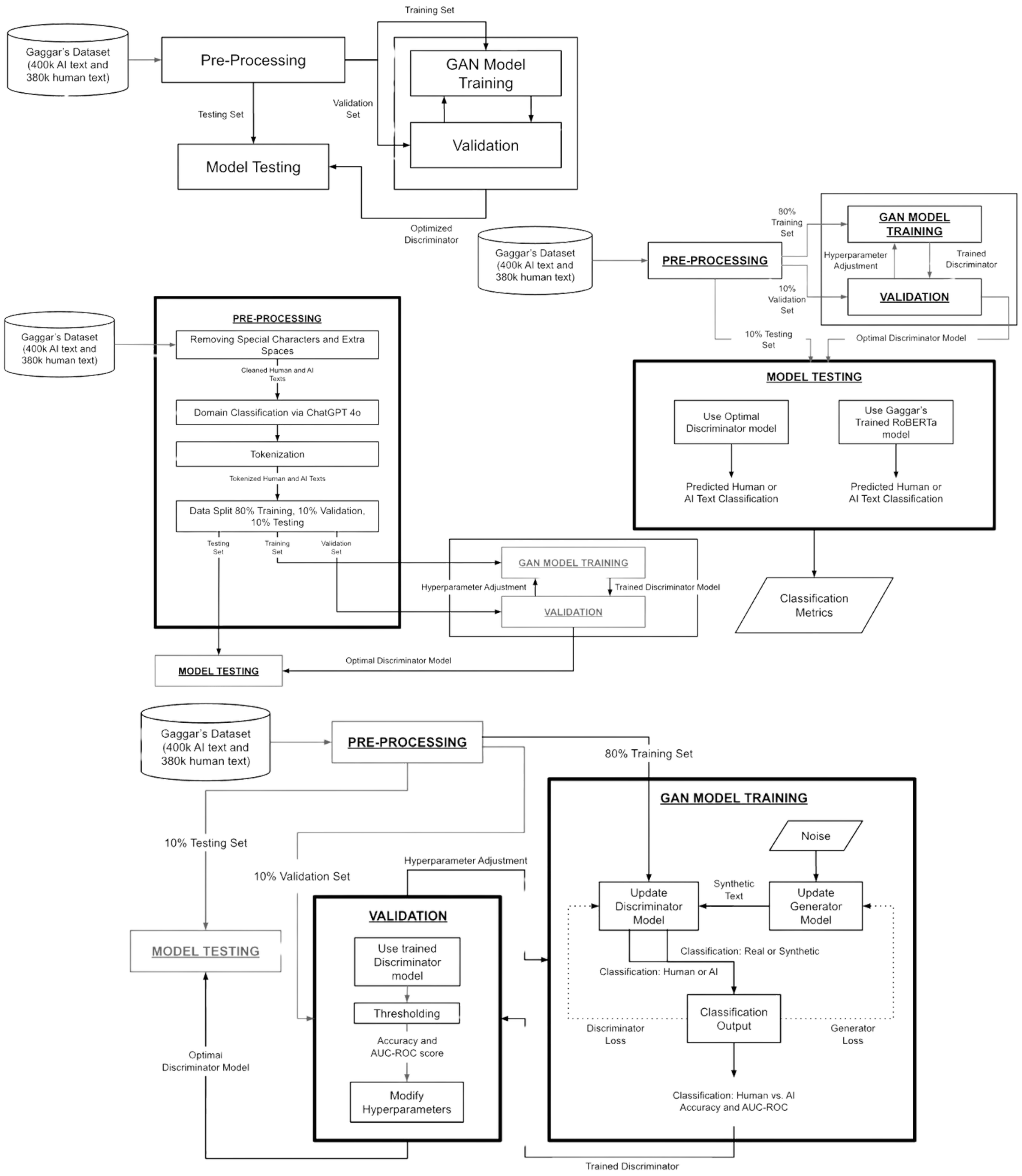
B. Background of the Study

This study builds on previous work by proposing the use of GANs to further enhance RoBERTa's capabilities, demonstrated in Gaggar et al. (2023), where RoBERTa already showed strong performance. By incorporating adversarial training and data augmentation with GANs, we aim to advance the current state of AI-generated text detection, pushing for higher AUC-ROC scores and robust model evaluations across diverse textual domains.

C. Statement of the Problem

1. How can the GAN model be configured to improve text classification performance for AI-generated text detection compared to the model of Gaggar et al.?
2. Will the use of GAN be able to achieve a higher AUC-ROC than what was achieved by Gaggar et al.?
3. How will the proposed GAN model perform on categorized AI-generated and human-written texts across different contexts?

D. System Architecture


Gaggar's Dataset (400k AI text and 380k human text)
Pre-Processing
Training Set
GAN Model Training
Validation Set
Validation
Testing Set
Model Testing
Optimized Discriminator
Gaggar's Dataset (400k AI text and 380k human text)
PRE-PROCESSING
80% Training Set
GAN MODEL TRAINING
Hyperparameter Adjustment
Trained Discriminator
10% Validation Set
VALIDATION
10% Testing Set
Optimal Discriminator Model
MODEL TESTING
Use Optimal Discriminator model
Use Gaggar's Trained RoBERTa model
Predicted Human or AI Text Classification
Predicted Human or AI Text Classification
Classification Metrics
Gaggar's Dataset (400k AI text and 380k human text)
PRE-PROCESSING
Removing Special Characters and Extra Spaces
Cleaned Human and AI Texts
Domain Classification via ChatGPT 4o
Tokenization
Tokenized Human and AI Texts
Data Split 80% Training, 10% Validation, 10% Testing
Testing Set
Training Set
Validation Set
GAN MODEL TRAINING
Hyperparameter Adjustment
Trained Discriminator Model
VALIDATION
Optimal Discriminator Model
MODEL TESTING
Gaggar's Dataset (400k AI text and 380k human text)
PRE-PROCESSING
80% Training Set
10% Testing Set
10% Validation Set
Hyperparameter Adjustment
MODEL TESTING
VALIDATION
Use trained Discriminator model
Thresholding
Accuracy and AUC-ROC score
Modify Hyperparameters
Optimal Discriminator Model
GAN MODEL TRAINING
Noise
Update Discriminator Model
Synthetic Text
Update Generator Model
Classification: Real or Synthetic
Classification: Human or AI
Classification Output
Discriminator Loss
Generator Loss
Classification: Human vs. AI Accuracy and AUC-ROC
Trained Discriminator

E. Objectives of the Study

To fine-tune the GAN model to enhance text classification performance and improve upon the model of Gaggar et al.
To achieve a higher AUC-ROC in classifying AI-generated text and human-written text compared to the 95.24% that was achieved by Gaggar et al.
To evaluate the comparative performance of the GAN model per AI-generated text and human-written text domain.

G. Conclusion

- RoBERTa-Base Superiority: Outperforms GAN-Discriminator with **higher AUC-ROC up to 97.31% versus 92.55%.**
- Statistical Significance: Confirmed superior performance of RoBERTa through non-overlapping confidence intervals.
- GAN Text Quality: Faces issues with coherence and quality; needs futher fine-tuning.
- Performance: GAN excels in structured domains like Sports and Education but lags in complex areas such as Health and Politics, indicating a need for domain-specific improvements.

F. Results and Discussion

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

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%

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**

- Overall Performance Metrics: The first table compares performance between human and AI texts, indicating **higher precision and recall for AI texts over human texts.**
- Model Comparison: The second table shows **AUC-ROC, training, validation, and test performance** for the original RoBERTa-base model, a replicated RoBERTa-base model, and a GAN-Discriminator. The replicated RoBERTa-base model exhibits improved performance on the test dataset, achieving a **99.31% AUC-ROC**, while the GAN-Discriminator shows strong validation and test performances, albeit slightly lower than the replicated RoBERTa model.
- Domain-Specific Performance: The third table details the models' **performance across various domains**, highlighting variations in AUC-ROC, accuracy, precision, recall, and F1-score. The model performs exceptionally well in the Sports domain but shows lower effectiveness in more nuanced domains like Health and Politics, suggesting varying levels of adaptability to different content types.