



REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

**ENSURING ACADEMIC INTEGRITY:
TRANSFORMER-BASED AND ENSEMBLE
MACHINE LEARNING APPROACHES FOR
DIFFERENTIATING HUMAN-WRITTEN AND AI-
GENERATED TEXT**

Layth Rafea HAZIM

Doctor of Philosophy

Supervisor

Assoc. Prof. Dr. Oğuz ATA

İstanbul, 2025

**ENSURING ACADEMIC INTEGRITY: TRANSFORMER-BASED
AND ENSEMBLE MACHINE LEARNING APPROACHES FOR
DIFFERENTIATING HUMAN-WRITTEN AND AI-GENERATED
TEXT**

Layth Rafea HAZIM

Department of Electrical and Computer Engineering

Doctor of Philosophy

ALTINBAŞ UNIVERSITY

2025

The dissertation titled ENSURING ACADEMIC INTEGRITY: TRANSFORMER-BASED AND ENSEMBLE MACHINE LEARNING APPROACHES FOR DIFFERENTIATING HUMAN-WRITTEN AND AI-GENERATED TEXT prepared by LAYTH RAFEHA HAZIM HAZIM and submitted on 12/12/2025 has been **accepted unanimously** for the degree of PhD in Electrical and Computer Engineering.

Assoc. Prof. Dr. Oğuz ATA

Supervisor

Dissertation Defense Committee Members:

Assoc. Prof. Dr. Oğuz ATA

Department of Computer
Engineering,

İstanbul Atlas University

Prof. Dr. Hasan Hüseyin BALIK

Department of Computer
Engineering,

İstanbul Atlas University

Assoc. Prof. Dr. Sefer KURNAZ

Department of Computer
Engineering,

Altınbaş University

Assoc. Prof. Dr. Aytuğ BOYACI

Department of Computer
Engineering,

National Defence University

Asst. Prof. Dr. Can EYÜPOĞLU

Department of Computer
Engineering,

National Defence University

I hereby declare that this dissertation meets all format and submission requirements for a PhD dissertation.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Layth Rafea HAZIM

Signature

XXXXXXXXXX

DEDICATION

All praises and thanks to the Almighty, Allah (SWT), who helps me to finish this study, Allah gives me the opportunity, strength and the ability to complete my study for a PHD degree after a long time of continuous work. No volume of words is enough to express my gratitude towards my supervisor, Dr. Oğuz Ata without his knowledge and assistance plus his recommendations this study would not have been successful, he has helped me explore this topic in an organized manner and provided me with all the ideas on how to work towards a research-oriented venture.

Finally, it would not be possible for me to complete the study and this project without the help of Allah and then supporting and encourage from my family and friends. First and foremost, my gratitude goes to my father, my mother and my wife to motivate me and for their endless support for me.

Thanks for all persons who helped or contributed to finish my PHD program.

PREFACE

This dissertation includes content that has previously appeared in peer-reviewed publications authored by the candidate:

Hazim, L. R., & Ata, O. (2024). Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for Text Classification. In Proceedings of the 2024 16th International Symposium on Electronics and Telecommunications (ISETC 2024), IEEE, Timișoara, Romania. <https://doi.org/10.1109/ISETC63109.2024.10797291>

Hazim, L. R., & Ata, O. (2025). Bridging the Gap: Ensemble Learning-Based NLP Framework for AI-Generated Text Identification in Academia. (Accepted for publication, SCIE-indexed journal).

Relevant material from these publications has been revised and integrated throughout the dissertation to ensure consistency with the doctoral thesis structure, formatting standards, and university guidelines. All reused content is the original work of the candidate and has been appropriately adapted for this comprehensive presentation. The inclusion of previously published material complies with the respective publishers' policies and upholds all academic integrity and copyright requirements.

ABSTRACT

ENSURING ACADEMIC INTEGRITY: TRANSFORMER-BASED AND ENSEMBLE MACHINE LEARNING APPROACHES FOR DIFFERENTIATING HUMAN-WRITTEN AND AI-GENERATED TEXT

HAZIM, Layth Rafea

Ph.D., Department of Electrical and Computer Engineering, Altınbaş University

Supervisor: Assoc. Prof. Dr. Oğuz ATA

Date: 12/2025

Pages: 138

The explosive development of large language models (LLMs) such as ChatGPT and Bard transformed text creation, offering new possibilities for academic communication and raising deep concerns regarding authenticity and authorship. As AI-created content has come to be almost indistinguishable from human writing, the need for open and reliable mechanisms to authenticate authorship has become crucial in sustaining academic trust.

This work presents and illustrates two mutually complementary Natural Language Processing (NLP) frameworks the Transformer-Based Detection Framework and the Ensemble Learning-Based Detection Framework and contrasts them on three evenly balanced datasets: AI-GA, HWAI, and HAGT-1M. The former uses Sentence-BERT (SBERT) and RoBERTa embeddings together with Logistic Regression (LR) and Feed-Forward Neural Networks (FNNs) to detect deep semantic and syntactic trends. Across datasets, transformer configurations achieved strong performance: SBERT+LR: 93.66% (AI-GA), 91.37% (HWAI), 99.64% (HAGT-1M); RoBERTa+FNN: 91.13% (AI-GA), 90.29% (HWAI), 99.95% (HAGT-1M) with RoBERTa-FNN peaking at 99.95% on HAGT-1M.

To ensure transparency and surmount the "black-box" limitation of transformers, the second method leverages ensemble learning by combining LR, SVM, Random Forest,

Extremely Randomized Trees, XGBoost, AdaBoost, and SGD with soft voting over TF-IDF and linguistic feature sets (n-grams, lexical diversity, sentence length). The ensemble achieved 99.88% (AI-GA), 99.37% (HWAI), and 100% (HAGT-1M), outperforming prior state-of-the-art and demonstrating increased robustness on varied academic text.

The current study focuses on reproducible preprocessing, data exploration, and careful threshold tuning regarding sensitivity and fairness, especially for non-native research scholars. Model linguistic authenticity as a three-dimensional property present in syntax, semantics, and stylistic coherence. Offer open, scalable, and transparent detection protocols suitable for universities, publishers, and research integrity offices.

Future research will include the investigation of hybrid quantum machine learning (HQML), explainable-AI tools with more detail (e.g., SHAP/LIME/attention visualization), and multilingual extension in order to maintain cross-domain generalization.

Keywords: AI-Generated Text Detection, Transformer Models, Ensemble Learning, Academic Integrity, Explainable Artificial Intelligence, Natural Language Processing, Hybrid Quantum Machine Learning.

TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACT	vii
LIST OF TABLES.....	xv
LIST OF FIGURES.....	xvi
ABBREVIATIONS.....	xx
LIST OF SYMBOLS.....	xxii
1. INTRODUCTION	1
1.1 BACKGROUND AND MOTIVATION	1
1.2 PROBLEM STATEMENT	4
1.3 RESEARCH QUESTIONS AND OBJECTIVES	5
1.3.1 Research Questions (RQs).....	5
1.3.2 Research Objectives (ROs).....	6
1.4 SIGNIFICANCE OF THE STUDY	6
1.5 THESIS STRUCTURE.....	7
2 LITERATURE REVIEW.....	9
2.1 INTRODUCTION	9
2.2 EVOLUTION OF AI-GENERATED TEXT AND NATURAL LANGUAGE PROCESSING	9
2.3 ETHICAL AND ACADEMIC CHALLENGES OF AI-GENERATED TEXT ..	11
2.4 HUMAN PERCEPTION AND COGNITIVE BIAS IN AI-GENERATED CONTENT RECOGNITION	13
2.5 MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR AI- GENERATED TEXT DETECTION.....	14

2.5.1	Traditional Machine Learning Approaches	14
2.5.2	Deep Learning and Transformer-Based Models	15
2.5.3	Ensemble Learning and Interpretability	15
2.5.4	Performance and Dataset Considerations	16
2.6	DATASETS, BENCHMARKING, AND EVALUATION STRATEGIES	16
2.6.1	Dataset Construction and Composition	17
2.6.2	Benchmarking and Evaluation Metrics	18
2.6.3	Reproducibility and Open-Source Practices	18
2.7	SUMMARY AND RESEARCH GAPS	19
3	METHODOLOGY	21
3.1	INTRODUCTION	21
3.2	SOFTWARE USED	22
3.3	DATASETS DESCRIPTION	23
3.3.1	AI-GA Dataset (Academic Abstracts)	23
3.3.2	HWAI Dataset (Human–AI Academic Writing Corpus)	25
3.3.3	HAGT-1M Dataset (Human–AI Generated Text Corpus)	27
3.4	SCENARIO 1: TRANSFORMER-BASED DETECTION FRAMEWORK	29
3.4.1	Overview	29
3.4.2	Text Representation	31
3.4.3	Classification Models	34
3.4.3.1	Logistic Regression (LR)	34
3.4.3.2	Feedforward Neural Network (FNN)	35

3.4.3.3	Training configuration	35
3.4.3.4	Summary	36
3.4.4	Evaluation and Performance Metrics	36
3.4.4.1	Performance metrics	36
3.4.4.2	Validation strategy	37
3.4.4.3	Computational environment	38
3.4.4.4	Summary	38
3.5	SCENARIO 2: ENSEMBLE LEARNING-BASED DETECTION FRAMEWORK	38
3.5.1	Overview	38
3.5.2	Linguistic and Statistical Feature Engineering.....	41
3.5.2.1	Exploratory Data Analysis (EDA).....	41
3.5.2.2	Linguistic Analysis (N-grams and Frequent Words).....	45
3.5.2.3	Statistical Feature Extraction (TF-IDF).....	51
3.5.3	Machine Learning Models.....	54
3.5.3.1	Logistic Regression (LR).....	54
3.5.3.2	Support Vector Machine (SVM).....	55
3.5.3.3	Random Forest (RF)	55
3.5.3.4	Extremely Randomized Trees (ErT).....	55
3.5.3.5	Stochastic Gradient Descent (SGD)	56
3.5.3.6	XGBoost	56
3.5.3.7	AdaBoost	56
3.5.4	Ensemble Model Construction	56
3.5.5	Evaluation and Statistical Validation	58

3.5.6	Summary of Scenario 2	59
3.6	ETHICAL CONSIDERATIONS.....	59
3.7	SUMMARY OF THE METHODOLOGY	60
4	RESULT AND ANALYSIS	61
4.1	INTRODUCTION	61
4.2	SCENARIO 1: TRANSFORMER-BASED FRAMEWORK RESULTS	62
4.2.1	Model Configuration	62
4.2.2	Results on AI-GA Dataset	63
4.2.3	Results on HWAI Dataset	67
4.2.4	Results on HAGT-1M Dataset	72
4.2.5	Comparative Analysis and Summary of Scenario 1	76
4.2.5.1	Model-level comparison	76
4.2.5.2	Embedding-level comparison	77
4.2.5.3	Dataset-level observations	77
4.2.5.4	Summary	77
4.3	SCENARIO 2: ENSEMBLE LEARNING-BASED FRAMEWORK RESULTS	78
4.3.1	Model Configuration	78
4.3.2	Model Training Performance.....	80
4.3.3	Results on AI-GA Dataset	80
4.3.4	Results on HWAI Dataset	83
4.3.5	Results on HAGT-1M Dataset	85
4.3.6	Comparative Analysis and Summary of Scenario 2	87
4.4	SUMMARY OF RESULTS AND ANALYSIS	89

5	DISCUSSION AND CONCLUSIONS	91
5.1	INTRODUCTION	91
5.2	DISCUSSION OF KEY FINDINGS	91
5.2.1	Transformer-Based Framework (Scenario 1)	91
5.2.2	Ensemble Learning-Based Framework (Scenario 2).....	92
5.2.3	Overall Insights	92
5.3	THEORETICAL AND PRACTICAL IMPLICATIONS	93
5.3.1	Theoretical Implications	93
5.3.2	Practical Implications	93
5.4	COMPARATIVE EVALUATION OF FRAMEWORKS	94
5.5	ETHICAL AND SOCIETAL IMPLICATIONS	94
5.6	COMPARISON WITH PREVIOUS WORK	95
5.6.1	Overview of Prior Approaches	95
5.6.2	Performance of the Proposed Frameworks.....	96
5.6.3	Discussion.....	97
5.7	LIMITATIONS AND FUTURE WORK	98
5.7.1	Limitations.....	98
5.7.2	Future Work.....	99
5.7.3	Summary.....	100
5.8	FINAL CONCLUSIONS.....	100
	REFERENCES	103
	APPENDIX A.....	112
	APPENDIX B.....	114

APPENDIX C..... 117



LIST OF TABLES

	<u>Pages</u>
Table 3.1: Frequent Words in AI and Human Text of (AI-GA) Dataset.....	50
Table 3.2: Frequent Words in AI and Human Text of (HWA) Dataset.	50
Table 3.3: Frequent Words in AI and Human Text of (HAGT-1M) Dataset.....	50
Table 4.1: Models Configuration Parameters.....	62
Table 4.2: Performance Metrics for AI-GA Dataset Using Cross-Validation.	63
Table 4.3: Performance Metrics for HWA Dataset Using Cross-Validation.....	68
Table 4.4: Performance Metrics for HAGT-1M Dataset Using Cross-Validation.....	72
Table 4.5: The Best Parameters Utilized for ML Models.	79
Table 4.6: Results for the Training Models of Datasets.....	80
Table 4.7: Results for the Testing Models of AI-GA Dataset.	81
Table 4.8: Results for the Testing Models of HWA Dataset.	83
Table 4.9: Results for the Testing Models of HAGT-1M Dataset.	85
Table 5.1: Comparative Performance of Transformer-Based and Ensemble Learning-Based Frameworks across Datasets.....	94
Table 5.2: Baseline Comparison of the Proposed Framework with Previous Studies.	96
Table B.1: Overview of Dataset Composition and Class Balance.	114

LIST OF FIGURES

	<u>Pages</u>
Figure 3.1: Conceptual Framework of the Proposed Methodology for AI-generated Academic Text Detection.	22
Figure 3.2: Interface of the Anaconda Navigator Environment Used in This Study.	23
Figure 3.3: Snapshot of the AI-GA Dataset Structure Showing the Fields: Title, Abstract, and Label.	24
Figure 3.4: Distribution of AI-generated and Human-written Abstracts in the AI-GA Dataset.	25
Figure 3.5: Sample Structure of the HWAI Dataset Showing the Text and Class Fields Representing Human and AI-generated Articles.	26
Figure 3.6: Distribution of Human-written and AI-generated Articles in the HWAI Dataset.	26
Figure 3.7: Sample Structure of the HAGT-1M Dataset Showing text Entries and Binary Labels.	28
Figure 3.8: Distribution of AI-generated and Human-written Text Samples in the HAGT-1M Dataset.	28
Figure 3.9: Framework of the Transformer-Based Detection Model.	29
Figure 3.10: Architecture of SBERT and RoBERTa for Generating Sentence Embeddings.	33
Figure 3.11: Architecture of SBERT and RoBERTa.	34
Figure 3.12: Ensemble Learning-Based Detection Framework.	39
Figure 3.13: Word Clouds for the AI-GA Dataset Combined.	42
Figure 3.14: Word Clouds for the AI-GA Dataset Human.	42

Figure 3.15: Word Clouds for the AI-GA Dataset AI.	42
Figure 3.16: Word Clouds for the HWAI Dataset Combined.	43
Figure 3.17: Word Clouds for the HWAI Dataset Human.	43
Figure 3.18: Word Clouds for the HWAI Dataset AI.	44
Figure 3.19: Word Clouds for the HAGT-1M Dataset Combined.	44
Figure 3.20: Word Clouds for the HAGT-1M Dataset Human.	44
Figure 3.21: Word Clouds for the HAGT-1M Dataset AI.	45
Figure 3.22: Frequency of 1-grams in AI-GA Dataset.	46
Figure 3.23: Frequency of 3-grams in AI-GA Dataset.	46
Figure 3.24: Frequency of 5-grams in AI-GA Dataset.	46
Figure 3.25: Frequency of 1-grams in HWAI Dataset.	47
Figure 3.26: Frequency of 3-grams in HWAI Dataset.	47
Figure 3.27: Frequency of 5-grams in HWAI Dataset.	48
Figure 3.28: Frequency of 1-grams in HAGT-1M Dataset.	48
Figure 3.29: Frequency of 3-grams in HAGT-1M Dataset.	49
Figure 3.30: Frequency of 5-grams in HAGT-1M Dataset.	49
Figure 3.31: TF-IDF Feature Extraction Workflow.	53
Figure 3.32: Ensemble Learning Architecture for AI-GA Dataset.	57
Figure 3.33: Ensemble Learning Architecture for HWAI Dataset.	58
Figure 4.1: Confusion Matrix for SBERT + LR on AI-GA Dataset.	64
Figure 4.2: ROC Curve for SBERT + LR on AI-GA Dataset.	64
Figure 4.3: Confusion Matrix for SBERT + FNN on AI-GA Dataset.	64

Figure 4.4: ROC Curve for SBERT + FNN on AI-GA Dataset.....	65
Figure 4.5: Accuracy Plot for SBERT + FNN on AI-GA Dataset.	65
Figure 4.6: Confusion Matrix for RoBERTa + LR on AI-GA Dataset.	66
Figure 4.7: ROC Curve for RoBERTa + LR on AI-GA Dataset.....	66
Figure 4.8: Confusion Matrix for RoBERTa + FNN on AI-GA Dataset.	66
Figure 4.9: ROC Curve for RoBERTa + FNN on AI-GA Dataset.....	67
Figure 4.10: Accuracy Plot for RoBERTa + FNN on AI-GA Dataset.	67
Figure 4.11: Confusion Matrix for SBERT + LR on HWAI Dataset.....	68
Figure 4.12: ROC Curve for SBERT + LR on HWAI Dataset.	69
Figure 4.13: Confusion Matrix for SBERT + FNN on HWAI Dataset.	69
Figure 4.14: ROC Curve for SBERT + FNN on HWAI Dataset.	69
Figure 4.15: Accuracy Plot for SBERT + FNN on HWAI Dataset.....	70
Figure 4.16: Confusion Matrix for RoBERTa + LR on HWAI Dataset.	70
Figure 4.17: ROC Curve for RoBERTa + LR on HWAI Dataset.	70
Figure 4.18: Confusion Matrix for RoBERTa + FNN on HWAI Dataset.....	71
Figure 4.19: ROC Curve for RoBERTa + FNN on HWAI Dataset.	71
Figure 4.20: Accuracy Plot for RoBERTa + FNN on HWAI Dataset.	71
Figure 4.21: Confusion Matrix for SBERT + LR on HAGT-1M Dataset.....	73
Figure 4.22: ROC Curve for SBERT + LR on HAGT-1M Dataset.	73
Figure 4.23: Confusion Matrix for for SBERT + FNN on HAGT-1M Dataset.	73
Figure 4.24: ROC Curve for for SBERT + FNN on HAGT-1M Dataset.....	74
Figure 4.25: Accuracy Plot for for SBERT + FNN on HAGT-1M Dataset.....	74

Figure 4.26: Confusion Matrix for RoBERTa + LR on HAGT-1M Dataset.	74
Figure 4.27: ROC Curve for RoBERTa + LR on HAGT-1M Dataset.	75
Figure 4.28: Confusion Matrix for RoBERTa + FNN on HAGT-1M Dataset.	75
Figure 4.29: ROC Curve for RoBERTa + FNN on HAGT-1M Dataset.	75
Figure 4.30: Accuracy Plot for for RoBERTa + FNN on HAGT-1M Dataset.	76
Figure 4.31: ROC Curve for Voting Classifier on AI-GA Dataset.	81
Figure 4.32: Confusion Matrix for Voting Classifier on AI-GA Dataset.	82
Figure 4.33: Class Prediction Error for Voting Classifier on AI-GA Dataset.	82
Figure 4.34: ROC Curve for Voting Classifier on HWAI Dataset.	83
Figure 4.35: Confusion Matrix for Voting Classifier on HWAI Dataset.	84
Figure 4.36: Class Prediction Error for Voting Classifier on HWAI Dataset.	84
Figure 4.37: ROC Curve for Voting Classifier on HAGT-1M Dataset.	86
Figure 4.38: Confusion Matrix for Voting Classifier on HAGT-1M Dataset.	86
Figure 4.39: Class Prediction Error for Voting Classifier on HAGT-1M Dataset.	86
Figure B.1: AI-GA Dataset Cross-Section — Sample AI vs. Human Abstracts.	115
Figure B.2: HWAI Dataset Cross-Section — Sample AI vs. Human Articles.	115
Figure B.3: HWAI Dataset Cross-Section — Sample AI vs. Human Articles.	116

ABBREVIATIONS

ACC	: Accuracy
AIGT	: AI-Generated Text
AI-GA	: Artificial Intelligence–Generated Academic Abstracts Dataset
ANN	: Artificial Neural Network
API	: Application Programming Interface
AUC	: Area Under the Curve
CC	: Cross-Correlation
CNN	: Convolutional Neural Network
CORD-19	: COVID-19 Open Research Dataset
EDA	: Exploratory Data Analysis
ErT	: Extremely Randomized Trees
FNN	: Feed-Forward Neural Network
FP	: False Positive
FN	: False Negative
HQML	: Hybrid Quantum Machine Learning
HWAI	: Human–AI Academic Writing Corpus
LLM	: Large Language Model
LR	: Logistic Regression
LSTM	: Long Short-Term Memory
ML	: Machine Learning
MGT	: Machine-Generated Text
NLP	: Natural Language Processing

OOB	: Out-of-Bag
PREC	: Precision
RF	: Random Forest
ROC	: Receiver Operating Characteristic
RoBERTa	: Robustly Optimized BERT Pretraining Approach
SBERT	: Sentence-Bidirectional Encoder Representations from Transformers
SGD	: Stochastic Gradient Descent
SHAP	: SHapley Additive exPlanations
SVM	: Support Vector Machine
TDMA	: Time Division Multiple Access
TF-IDF	: Term Frequency–Inverse Document Frequency
TN	: True Negative
TP	: True Positive
XGB	: Extreme Gradient Boosting (XGBoost)

LIST OF SYMBOLS

μ	: Mean or Mobility (context-dependent)
λ	: Wavelength
ω	: Angular Frequency
α	: Learning Rate / Regularization Coefficient
β	: Model Weight / Coefficient
σ	: Standard Deviation
ε	: Error Term
θ	: Model Parameter Vector
ρ	: Correlation Coefficient
η	: Efficiency / Step Size (in SGD)
τ	: Threshold Value
π	: Mathematical Constant (≈ 3.1416)

1. INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

In the last ten years, AI or Artificial Intelligence has developed a lot, from being a niche computational field to becoming a major force that changes the whole area of human knowledge [1]. One of the areas where AI has made a remarkable difference is Natural Language Processing (NLP), where systems like OpenAI's ChatGPT and Google's Bard have become the main players and have really influenced the production, interpretation, and distribution of text information in a whole new way [2]. LLMs (Large Language Models) based on the transformer architecture have performed exceptionally in the different applications of machine translation, automatic answering of questions, summarizing, content and academic writing, etc., and have managed to create a human-like language across all these areas [3].

Vaswani et al. (2017) [4], have introduced the transformer that is at the core of these models and this has led to major developments in language representation with the help of its novel self-attention mechanism through which models could capture not only the dependencies but also the relationships among the semantics over long spans of text without much difficulty as compared to previous RNNs or CNNs. This milestone opened the door to the development of architecture that is deep and bidirectional such as BERT, RoBERTa, and Sentence-BERT (SBERT), which today's generative models are primarily based on [5]. Their ability to learn embeddings that are sensitive to context from huge text collections has made it difficult to distinguish artificial and human communication, and this has led to the formation of AI systems that can produce text that is similar to that of humans [6].

By unlocking new paths for both automation and creative use of language, these developments have raised some serious ethical, academic, and epistemological issues at the same time. When AI can write essays, paper, and articles, it forces us to rethink who is the author, what is original, and who is responsible for intellectual rights [7]. In the academic setting where the authenticity and source of knowledge are very important, the presence of AI-generated text (AIGT) may lead to loss of integrity in research and also the peer-review process [8]. The submission of AI-written abstracts or the using of fake citations has

caused several universities, publishers, and conferences to become active in re-evaluating their authorship policies and creating new disclosure rules [9]. To battle the power of AI tools in manuscript preparation, the Association for Computational Linguistics (ACL) and the International Conference on Machine Learning (ICML) have been two of the major academic bodies that have recently changed their authorship criteria formally to accommodate this trend [10].

The societal consequences that the world faces are not only limited to academia. AI's content generation application has been a double-edged sword for professionals dealing with information, writing, and managing due to the issue it created around misinformation, and the problem of plagiarism, and loss of creativity [11]. On the other hand, the systems that facilitate adaptive tutoring and multilingual access could at the same time lead to the large-scale creation of misinformation or biased content [12]. This duality of innovation and malpractice underlines the need for safeguards that authenticate the text and ensure that the AI tool is an assistant, not a deceptive replacement for human minds [13].

From a computational viewpoint, the detection of text produced by AI is not an easy problem to solve. The initial approaches used stylometric features that were on the surface level such as frequency of words, sentence length, and syntactic structures, allowing AI to produce text which was readable or redundant in terms of phrasing [14]. But with the progressive advancement of the transformer model, its ability to manipulate and keep coherence with the context increased dramatically minimizing the efficacy of such techniques [15]. Deep learning discriminators are so powerful, but their operational procedure is typically in such a way that it cannot be understood easily thus preventing their interpretability and reproducibility [16]. This black-box nature of their operation is a trust-reducing factor especially in academia where the need for explainability and accountability is paramount [17].

The academia has reacted to this scenario in the past few years by flooding the market with research on AI authorship detection that employs semantic embeddings as well as ensemble learning methods [18]. Still the situation is not ideal; there are a lot of studies that deal predominantly with general-purpose datasets like social media posts or news articles, instead of scholarly writing which bears different lexical and rhetorical structures [19]. Furthermore, reproducible open-source datasets that accurately mirror real academic

writing styles are still not widely available and thus objective benchmarking and cross-study comparison are hindered [20].

Underlying this scenario is the rising demand from three different fronts which together form the driving force for this investigation.

- a. Theoretical need: to get the data-driven understanding of the differences in the linguistic and semantic areas between the machine-generated and human-written academic texts.
- b. Technical need: to create the interpretable, scalable frameworks that integrate deep contextual embeddings and transparent machine-learning models for the precise authorship detection.
- c. Ethical need: to take the part in the effort that protects academic integrity by supplying the reproducible methods that help with fair and non-discriminatory detection without penalizing the legitimate human diversity in writing style.

Therefore, this project plans to create the all-encompassing dual-scenario framework that combines all these goals. The first scenario sets the standard using the transformer-based embeddings Sentence-BERT (SBERT) and RoBERTa combined with classifiers like Logistic Regression and Feedforward Neural Networks to check the validity of the capture of the authorship's stylistic cues by the contextual representations. The second scenario takes this solid ground further by applying ensemble learning that is further supported by linguistic and statistical features (TF-IDF, n-grams, lexical diversity) to improve transparency and generalization across different academic corpora.

The proponents of the two approaches, when used together, bring to the center of argument the detection of AI authorship, thus beginning the journey from deep semantic modeling to explainable machine learning. The analysis of the three well-proportioned datasets AI-GA, HWAI, and HAGT-1M is expected by this study to offer a replicable benchmark for subsequent work and also to maintain the ethical and methodological standards required to continue the trust in the era of artificial intelligence [21].

Ultimately, the extensive use of AI-written academic papers indicates not only the practicability of technology in such works but also the creation of hurdles in terms of ethics, which is why it is necessary to have the authorship verification methods that are capable of functioning efficiently. Studies have previously pointed out the differences in style and semantics of AI and human text but at the same time there are still significant gaps regarding explainability, dataset diversity, and cross-domain generalization. To address these challenges, the present research proposes an all-inclusive detection framework that merges deep contextual modeling with interpretable linguistic and ensemble learning techniques. The next section sets forth the problem statement, research objectives, and questions that will lead this investigation in the direction of a transparent and reproducible solution for separating AI-generated from human-written scholarly material.

1.2 PROBLEM STATEMENT

Current advances in Large Language Models (LLMs) and Natural Language Generation (NLG) GPT-4, Bard, and Claude have revolutionized text generation to produce very coherent, grammar-free, and contextually rich language [11]. These advances have made AI writing assistants, autocomplete tools, and research draft systems, speeding up automation in research and content creation. But this rapid growth poses significant societal and ethical issues, including misinformation, author manipulation, and intellectual responsibility loss [7].

One of the new challenges in academic publishing is the possibility of misuse of AI platforms to generate or assist in generating scholarly text. Compared with broad domains such as social media or news coverage, scientific writing is expected to advance novel information, meet discipline-specific norms, and be rigorously examined. The growing skill of AI systems to replicate these attributes makes it more probable that content created by AI will be mistaken for that produced by humans, which, in the long run, could harm the credibility of research results and the whole peer-review system. Important conferences like the 61st Annual Meeting of the Association for Computational Linguistics (ACL, 2023) and the Fortieth International Conference on Machine Learning (ICML, 2023) which identified these risks, have updated their author and disclosure policies in order to cover the ethical use of AI in scientific publications [9] [10].

While these issues have been raised, the current detection approaches are still inadequate in the academic context. Most of them are built for general text classification tasks and therefore cannot detect the differences in language, structure, and style that separate AI-generated academic text from real scholarly writing. The majority of transformer-based detection models also function as black boxes and their decisions do not provide much interpretability or reproducibility. This opacity makes it difficult to build trust and facilitate the use of these methods in formal academic settings.

A debatable point is that there is a huge requirement for an explainable and a comprehensive framework that can efficiently differentiate between the two types of texts, one being AI-written and the other human-written academic text. The current study aims to fulfil that requirement by implementing a twin-framework detection system that is based on transformer-based embeddings and ensemble learning coupled with statistical and linguistic analysis which provides both interpretability and high performance. The main objective is to sustain research integrity and also to encourage the responsible use of AI in academia.

1.3 RESEARCH QUESTIONS AND OBJECTIVES

1.3.1 Research Questions (RQs)

This research examines the issues raised by the Problem Statement and accordingly formulates the following fundamental research questions:

RQ1: Which structural and linguistic characteristics set apart academic texts produced by AI language models from those written by expert authors [22]?

RQ2: In what ways can exploratory data analysis (EDA) and linguistic feature extraction uncover the stylistic differences that exist between AI-academic writing and human texts [23]?

RQ3: How significantly do transformer-based embeddings (such as SBERT, RoBERTa) and ensemble learning models contribute to the accuracy and interpretability of AI-generated text identification in academic settings [24]?

RQ4: What is the comparison between hybrid architectures that merge deep embeddings and statistical models with those that depend solely on transformer-based or traditional

machine learning classifiers in terms of performance consistency, cross-domain generalizability, and interpretability [25]?

RQ5: What are the larger ethical and intellectual implications of implementing automated detection technologies for AI-generated academic works within the research community [26]?

1.3.2 Research Objectives (ROs)

The study is based on the following aims from which the questions arise:

RO1: To perform a comprehensive linguistic and exploratory analysis of both AI-generated and human-generated academic text, where their differences will be determined through the identification of discriminative features in coherence, lexical diversity, and syntactic structure.

RO2: To come up with and validate transformer-based detection models that use SBERT and RoBERTa embeddings to distinguish AI-generated research texts from human authorship.

RO3: To build a complementary ensemble learning framework that incorporates TF-IDF and n-gram features and multiple supervised classifiers (SVM, LR, RF, AdaBoost, XGBoost) to boost interpretability and generalizability.

RO4: To carry out comparative experiments on several datasets (AI-GA, HWAI, and HAGT-1M) to measure the performance of each model in terms of accuracy, precision, recall, F1-score, and ROC-AUC.

RO5: To hold a discussion and perform an exploration of the ethical, philosophical, and practical consequences of discerning between AI and human authored scholarly content, and to offer suggestions on research integrity and transparency in the area of scholarly publishing.

1.4 SIGNIFICANCE OF THE STUDY

The interdisciplinary nature of the research and its contribution to the domains of Natural Language Processing (NLP), Machine Learning (ML), and academic integrity are the main features of the study. As a result of the widespread use of Large Language Models (LLMs)

like GPT-4 and Bard, it has become very hard to tell whether a piece of writing is done by a human or an AI [18]. Such a situation puts the integrity and trustworthiness of academic publishing, peer review, and knowledge outlet channels at risk directly and thus endangers the whole process of research extraction, discussion, and dissemination [27].

The presentation of the two-framework architecture of a transformer-based detection system and an ensemble learning-based framework is scientifically very beneficial as it advances the methodological frontier of AI-authorship detection. The study integrates contextual embeddings (e.g., SBERT, RoBERTa) with interpretable statistical and linguistic features (e.g., TF-IDF, n-grams, lexical metrics), thus providing a comprehensive and reproducible solution to the problem of differentiating between AI-generated and human-written academic texts [28]. The findings are in line with the ongoing research in the area of explainable AI (XAI) and contribute to it by demonstrating the existence of a trade-off between interpretability and deep model complexity in the case of authorship verification [29].

The study, while ethical and practical, gives primary importance to academic integrity and transparency. If publishers, institutions, and educators want to have the manuscripts evaluated, the possible use of AI writing software to be disclosed, and the implementation of fair authorship policies, the proposed method can be their very foundation, [30]. The release of the balanced and open-source datasets (AI-GA, HWAI, HAGT-1M) provides the research community with a valuable benchmarking tool that facilitates reproducibility and collaboration in future studies, [31].

Lastly, this work has not just surpassed the time and social issue but also has been integral to the slow acceptance of responsible AI among scholars. Its value is not just in the performance parameters but also in ensuring that AI is part of the process that keeps the integrity of scientific argument and not the reason for loss of integrity [2].

1.5 THESIS STRUCTURE

The doctoral dissertation is composed of five major chapters, each focusing on a particular aspect of the inquiry and all together forming a detailed framework for distinguishing between AI-generated scholarly texts and those written by humans.

Ch 1: Introduction: In this chapter, the research background, motivation, and AI-generated academic work ethics are discussed. It raises the research problem, formulates the aims and questions, and indicates the theoretical, technological, and ethical significance of the study.

Ch 2: Literature Review: The study reviews thoroughly the literature on Natural Language Generation (NLG), authorship verification, and text classification. It points to the growth of machine learning and deep learning techniques in AI-authorship verification, uncovers the weaknesses of the current literature, and sets the conceptual groundwork for the proposed hypothesized dual-framework approach.

Ch 3: Methodology: The chapter starts with the research design and its setup. It explains the building and the initial processing of the dataset, the setup and tuning of the transformer-based model of both frameworks, namely the one based on ensemble learning and the one based on the single neural network, and the specific measures of performance which are reported. The part also places a strong emphasis on reproducibility by specifying the validation methods and the handling of ethical concerns.

Ch 4: Results and Analysis: This chapter is dedicated to the presentation of the experimental results and their interpretation for each of the two experiments. It is composed of the performance tables, confusion matrices, ROC curves, and visualizations of comparisons. The analysis is carried out in terms of accuracy, precision, recall, F1-score, and ROC-AUC for the different datasets (AI-GA, HWAI, and HAGT-1M), and then the discussion continues with trends and the efficiency of the framework.

Ch 5: Discussion and Conclusions: In the final chapter, the results are linked with research questions and the broader literature. The technical, ethical, and academic implications of the findings are analyzed with particular emphasis on their impact on authorship verification and academic integrity. The chapter wraps up with a presentation of the research results, limitations, and future studies to be conducted on hybrid and quantum-enhanced AI-detection models.

2 LITERATURE REVIEW

2.1 INTRODUCTION

The aim of this chapter is to present a detailed summary of the existing knowledge concerning the recognition of Artificial Intelligence – Generated Text (AIGT) specifically in the areas of academics and science. The review amalgamates the theoretical underpinnings, technological innovations, and methodological approaches from prior studies while pin-pointing the research gaps that the present work intends to fill.

The chapter, first of all, sketches the journey of AI text-creation and its union with the contemporary Natural Language Processing (NLP). Next, it points out the ethical and academic issues which are a direct result of the widespread use of large language models (LLMs) in the area of academic communication. After that, it reviews the empirical studies about people's perception of the AI-generated content, followed by a discussion of the machine-based detection methods, such as traditional ML, transformer-based and hybrid ensemble models. The penultimate part collects the revealed research gaps and situates the present study within the continuous academic conversation.

2.2 EVOLUTION OF AI-GENERATED TEXT AND NATURAL LANGUAGE PROCESSING

Symbolic music and rule-based visual art were the main forms of computer-generated content in the early days, and the outputs were such that the audiences could easily tell that they were not made by humans [32],[33]. The introduction of modern generative modeling brought about a rapid increase in the realism of the outputs: Generative Adversarial Networks (GANs) allowed for data-driven synthesis of images and texts, while the recently developed diffusion models went beyond GANs in quality metrics, setting a new record for content generation that is controllable [34],[35]. On the other hand, language technologies were being developed gradually from count-based n-gram pipelines to distributed representations and neural sequence models, which eventually led to fluency and coherence improvements in machine-produced texts [36],[37],[38].

Vastly powerful transformer architectures and their pretrained model ecosystem marked the turning point. The transformers not only acted as state-of-the-art language encoders but

also were the backbone of knowledge-driven tasks, such as treating language models as emergent knowledge bases and permitting pretraining through the sequence-to-sequence approach for generation, translation, and comprehension [39],[29]. Further gradual adaptation of LLMs to specific fields indicated both the potential and the limitations: Galactica—scientific models—showed reasoning over corpus of journals while at the same time exposing the risk of hallucination and bias in academic writing through language [40].

Initially offering public-facing LLMs, the large-scale AI text generation became necessitated. ChatGPT and Bard/Gemini, among others, were the systems that created the rapid and wide transition between the text written by humans and the one done by machines, making it available for the sectors of education, journalism, research, and many more for drafting, summarization, and creation of ideas [3],[2]. Only the study of the influence of AI-generated characters and agents on learning and well-being pointed to the technology's broad impact across society, though AI's presence had become substantial through practicality first [33]. This dual-path of capability growth and widespread usage has made the difference between human and machine authors less clear.

Better realism led to a shift of research priorities from stylistic fingerprints to contextual semantics. Classical stylometry (lexical frequencies, sentence length, surface syntax) was found inadequate versus the fluency of a transformer and thus detectors relied more on pretrained encoders (like SBERT and RoBERTa) and ensemble methods to capture higher-order semantics and discourse structure [37],[41]. Alongside, reproducible pipelines stressed preprocessing rigor (token normalization, stop word handling, and noise reduction) to stabilize downstream modeling and enable fair benchmarking across corpora [23],[36].

The literature points out that there are still challenges persisting over time, and these are surely very good challenges. Surveys and experts made a claim according to which detector robustness is likely to be shrinking in out-of-domain settings, human judges are usually very but not quite accurate in telling AI from human text, and there is necessity for developing evaluation protocols that keep in mind the backend evolution of the model families, prompts, and post-editing practices [41],[17],[42]. All of these trends taken together give rise to the need for the frameworks that will (i) fuse semantic embeddings

and interpretable feature sets, (ii) work with academic-oriented rather than just with general-domain corpora as the only criterion for evaluation, and (iii) the present study being one of those very priorities that rightly point out the issues of reproducibility and fair comparison that need to be addressed directly.

2.3 ETHICAL AND ACADEMIC CHALLENGES OF AI-GENERATED TEXT

The advent of Large Language Models (LLMs) has opened up a whole can of worms in terms of ethics, academic action, and society's perception of machine-generated text in terms of its authenticity and accountability. While AI software like ChatGPT and Bard are getting along famously with generating coherent and stylistically better texts, it is more and more challenging to distinguish the algorithmic creation from the human mind [3],[14]. Not only do these models present an array of compelling benefits for education, research, and communication; however, they also pose a package of risks, misinformation, plagiarism, and authorship manipulation [6],[32],[43], to mention but a few.

In the academic community, these are critically necessary boundaries. The academic community relies on transparency of authorship, verifiability of evidence, and novel contribution. The invasion of AI-generated scholarly activity, either by writing software or completed papers, is an affront to research validity and to peer-review systems' credibility [9],[10]. The case of gibberish or fake research papers produced by software such as SCIGen is a stark reminder of the capability of artificial content to slip through the editorial sieve and end up in the pages of reputable journals [40],[44]. Several other studies have shown that computer-generated texts still find their way in order to course through scientific databases, a testament to how the publication process is open to manipulation using generators [41].

The large-scale use of AI instruments by different industries, from an ethical point of view, has made it difficult to understand the concepts of intellectual property and putting one's cards on the table responsibly. The dilemma that institutions are facing today is that of innovation vs accountability where they are accountability innovators that help people who use AI instruments while at the same time the misuse of the instrument for deceiving is not allowed. conferences and publishers have responded to this tension by making policy changes. The author and submission policies of both the Association for Computational

Linguistics (ACL 2023) and the International Conference on Machine Learning (ICML 2023), for instance, have been explicitly revised to impose a requirement of disclosing AI's role in preparing a manuscript [9],[10]. These policies are the first steps towards establishing moral frontiers for generative tech in scientific discourse that are supported by institutions.

The impact of AI content generation is felt across different areas such as journalism, marketing, and education outside the academic world and this further exposure to fraud risks in information and consequently mistrust in the public [33],[43]. Even though generative models may have made the content creators' life easier, they are also the main reason behind the spread of persuasive misinformation on a large scale. Reader surveys suggest that most people have a stronger belief in AI's capability to produce high-quality news and opinion pieces than the same capabilities being applied to human counterparts. This is a paradoxical state in which the validity of content-generating algorithms becomes the source of believability rather than skepticism [11],[42]. Hence, automated text generation has been seen to be a double-edged sword; one side offers a breakthrough regarding technology while the other side offers an epistemic threat hence the need for detection, transparency, and literacy mechanisms that work effectively.

One more vital point is the issue of bias and fairness. The gigantic dataset that LLMs use for training contains a lot of web pages, and as a result, they also absorb and amplify the biases that were present in the original data sources such as social, cultural, and academic biases [39],[41]. If these models would be used for generating or summarizing academic texts, then they could very well be the ones that inadvertently misrepresent the voices of the marginalized or support false equivalences. Hence, the ethical AI writing practices demand not only the detection of synthetic content but also the interpretability and accountability in how AI systems are designed and applied.

The human aspect of the issue remains to be the core matter around which the discussions revolve. As AI systems become more and more involved with the writing process, the question comes up, after all, who is the author and to what extent can one get help that is considered acceptable? Should we treat AI as a co-author, a tool, or just a computational helper? Although the majority of academic institutions are against the idea of machines being authors, the lack of clear global standards leads to different and sometimes

contradicting practices. The existence of these conceptual and procedural gaps must be addressed to prevent the erosion of scholarly credibility and the loss of public trust in the era of generative intelligence.

2.4 HUMAN PERCEPTION AND COGNITIVE BIAS IN AI-GENERATED CONTENT RECOGNITION

The continuous and fast advancement of AI language models has rendered human detection of real and fake texts more and more unreliable. Older investigations indicated that people generally have a much higher opinion of their own abilities to spot AI-generated material, which is the case for most perceptual biases and inconsistencies [45],[17]. The surface-level cues such as syntax and tone are no longer sufficient for reliable differentiation because generative models like GPT-3, GPT-4, and Bard can create text that is almost indistinguishable from human writing. Therefore, researchers have changed their attention to the understanding of the impact of cognitive heuristics and familiarity effects on human judgment in the evaluation of machine-written content [42].

It is a fact established by evidence that readers often confuse the authorship of the text--- considering the AI-generated part as the human one, especially when such texts are presented in a credible or professional way [12]. In one experiment, the participants rated the synthetic scientific abstracts and news articles as more trustworthy than the genuine ones, which is a reflection of the credibility inversion phenomenon in which algorithmic fluency boosts perceived reliability [11]. This result is in line with previous behavioral research that has demonstrated that people depend on the very same stylistic fluency and grammatical consistency traits which LLMs now replicate with outstanding proficiency to draw conclusions about authenticity [17].

Detection ability has been shown to vary depending on the context according to research. For example, in the case of medical writing, academic essays, and creative literature, the participants exhibited lower accuracy rates than in the case of informal communication, which may indicate that human judging is affected by the level of disciplinary style familiarity [45]. Besides, the use of AI tools seems to cause a normalization effect, making the users pay less attention to the machine-generated content and thus less skeptical of it. The said trend could eventually lead to the blurring of the already-faint boundary between

human and artificial linguistic expression if the LLM-based assistants are used all the time as the writers' assistants.

Another side of the issue is that of the ethical framing and transparency. Surveys show that when the audience is made aware of the AI authorship of the text, at the same time the text is inconsistent with its factual content, and stylistic identity, their trust, and participation levels drop [42]. This psychological imbalance places a special emphasis on understanding the source and ethical disclosure as significant factors in the confidence issue of the user. The steps against misinformation cannot, thus, rely on technology-based detection systems but rather a combination of the same with pedagogical steps aimed at making individuals critical readers and digitally savvy.

To put it briefly, the way humans perceive AI-written content is limited by their mental shortcuts and the influence of the social and technological environment. The more advanced LLMs become in terms of language skills, the more difficult it will be to depend on human intuition as the only means of recognition. Hence, the best detection and awareness systems will have to meld together algorithmic classification with comprehension of human cognition, recognizing that in the battle for information integrity the human rater might gradually turn out to be the weakest link instead of the machine.

2.5 MACHINE LEARNING AND DEEP LEARNING APPROACHES FOR AI-GENERATED TEXT DETECTION

The detection of AI-generated content has passed through various methodological phases that started with the use of traditional machine learning (ML) models and subsequently progressed to deep learning (DL) and transformer-based architectures. Initially, the ML techniques identified linguistic discrepancies in the text generated automatically, whereas the modern approaches combine semantic embeddings, ensemble learning, and attention mechanisms to gain higher accuracy and interpretation [10],[32],[37].

2.5.1 Traditional Machine Learning Approaches

The first attempts to tell apart human-written texts from machine-generated ones were based on using the well-known classifiers such as Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forests (RF) [46],[47],[36].

Stylometric and statistical features such as word frequency, sentence length, and punctuation variance were utilized by these algorithms to pinpoint the anomalies in the text. The research conducted by Singh et al. [48] and Kowalczyk et al. [13] showed that even the quite simple models could obtain an accuracy level of over 90% on domain-specific datasets, for example, on Wikipedia or online review corpora. Nevertheless, these models often had a hard time crossing over to various writing domains because of their reliance on superficial linguistic cues rather than deeper contextual understanding [37].

2.5.2 Deep Learning and Transformer-Based Models

Deep neural networks (DNNs) together with transformer architectures fundamentally changed detection research. BERT, RoBERTa, and Sentence-BERT (SBERT) types of models made contextual perception more accurate as they could detect the relations and dependencies of the whole text and its parts [49],[29],[47]. Zellers et al. [6] developed Grover which is a generative-discriminative model that can produce and detect fake news articles with an accuracy of more than 92% [50]. Consistently, the studies demonstrating the application of BERT-based models to academic and statistical datasets reported F1-scores improvements over classical methods, thus confirming that contextual embeddings surpass the traditional feature extraction [51],[5].

Moreover, hands-on tests done in recent times have pointed out the importance of hybrid models where, for instance, CNN or RNN layers are combined with the optimization techniques or the ensemble methodologies. Oketunji [24] suggested a combination of CNN and RNN with over 92% accuracy in detection, meanwhile, Katib et al. [47] put forward the TSA-LSTM-RNN method which is associated with metaheuristic optimization, being able to achieve accuracy levels of more than 93%. This kind of mix gives high generalization and durability against the change of style in AI-produced text.

2.5.3 Ensemble Learning and Interpretability

Ensemble learning has become a widely accepted method for achieving the best results by combining the different strengths of models, and at the same time the models will be easier to understand and to protect against the problem of overfitting. Various methods like bagging, boosting, and soft voting are being used to predict the outcomes of the different classifiers and so make the whole process more reliable for diverse datasets [47],[52].

Alamleh et al. [7] and Ruopeng An et al. [53] have pointed out that not only do model ensembles lead to higher accuracy, but they also create a situation of stability when facing different linguistic patterns in academic and social media texts.

The major concern in the recent literature is the focus on explainable AI (XAI) as the key factor for detecting texts that are trustworthy and untrustworthy. By employing SHAP and LIME interpretability frameworks, researchers such as Kowalczyk et al. [13] and Pendyala et al. [54] were able to identify the most text features that weigh heavily on the decision of the model. These tools not only allow the model to be more transparent but also make it a human reviewer's responsibility to understand the model's reasoning. This is an important necessity for the ethical use of AI detectors in academia and journalism.

2.5.4 Performance and Dataset Considerations

Performance measurement is of highest importance in comparative research. Quality of detection is evaluated via metrics such as accuracy, precision, recall, F1-score, and AUC which are the gold standards [55],[56]. Although transformer-based and hybrid models exhibit optimal performance on benchmark data, their usual scalability and replicability still depend a great deal on dataset balance and quality [46]. Domain-specific corpora like the AI-GA, HWAI, and HAGT-1M prove that the depth of context and language plays a major role in the stability of models, thus the production of open-source datasets for fair benchmarking [52],[48],[24].

The overall picture drawn by the literature is that of a push towards hybrid and ensemble models that are capable of making use of linguistic and contextual features. However, interpretability, computational efficiency, and cross-domain generalization are still open challenges. Breaching these gaps would necessitate the use of frameworks that not only deliver high predictive performance but also adhere to the principles of transparency and ethical accountability that are central to the present research.

2.6 DATASETS, BENCHMARKING, AND EVALUATION STRATEGIES

The reliability and generalizability of AI-generated text detection systems heavily depend on the quality, diversity, and representativeness of the datasets that are used for model training and evaluation. The research community has been focusing more and more on the

construction of domain-specific and large datasets that are able to mirror the real-world variations in authorship, style, and intent [5],[48],[46], during the past few years.

2.6.1 Dataset Construction and Composition

The early researchers or professionals who worked on AI-text detection continued with the use of small and unbalanced corpora that were lacking in variety and in-depth analysis, and these often consisted of nothing more than short news articles, essays, or artificial paragraphs produced by GPT-2 or GPT-3 models [6],[13]. The limitations of the datasets indicated that they were unfit for validating the linguistic and rhetorical depth of academic writing, but they still provided support for initial testing. In an effort to conquer this challenge, a number of scientists have created specialized datasets for scientific and educational writing, capturing the subtle differences between human and AI authorship.

Salim and Hossain [5] came up with an Applied Statistics question answer dataset that mixes AI-generated and human-written answers to assess the context integrity of machine-produced scholarly material. In a similar fashion, Singh et al. [48] put together a Wikipedia–ChatGPT corpus containing 44,162 balanced articles that function as a benchmark for differentiating between factual encyclopedic writing and generated prose. It was found that classical models such as Logistic Regression could still achieve a remarkable 97% accuracy rate when trained on well-curated and balanced data, thereby emphasizing the significance of dataset quality over model complexity.

On the other hand, huge datasets such as HAGT-1M (Human and AI-Generated Text – 1 Million Samples) introduced by Oketunji [24] provide a very strong statistical power and cover many different domains. The HAGT-1M dataset, which has an equal number of human-written and AI-generated samples (500,000 for each), allows for model evaluation under the same and reproducible conditions. In the same vein, AI-GA and HWAI datasets representing academic and semi-academic writing, respectively, allow for the testing of domain adaptation and the evaluation of the generalization potential across different text types.

2.6.2 Benchmarking and Evaluation Metrics

In the literature, the evaluation of models has always been done using standard performance measures such as accuracy, precision, recall, F1-score, and ROC-AUC [55],[56],[46]. These metrics give a good picture of the classification's reliability and the trade-off between sensitivity (the correct identification of positives) and specificity (the correct rejection of negatives). But with the increasing intricacy of language models, perplexity and entropy-based measures were also used by the researchers to analyze and quantify the linguistic predictability and coherence [48].

For instance, the recent research that focused on the evaluation of AI-generated abstracts from PubMed utilized the perplexity-based analysis technique, which has indicated that the human-written abstracts are generally more unpredictable in terms of word sequence structure than the machine-generated ones [48]. The above-mentioned results strongly support the idea of combining traditional metrics with probabilistic ones to the fullest extent possible and thus, bringing about the proposed change of comprehensiveness in the evaluation process.

Moreover, cross-validation procedures especially 10-fold and stratified validation have turned into the norm for overfitting prevention and model certitude [47]. Researches with ensemble learning frameworks have published maximum robustness [52],[24]. Thus, it has been proved that the unification of different classes of classifiers can uplift the performance as well as the clarity of the model.

2.6.3 Reproducibility and Open-Source Practices

Reproducibility has become a main principle in the research work of AI-text detection. Nowadays many studies are supporting the need for transparent dataset documentation, code sharing and evaluation pipelines available through open repositories [55],[46]. The creation of open-source benchmarks not only allows for fair comparison of detection frameworks but also opens up the possibility of tracking the models' progress over time. However, there are still some datasets that are either proprietary or have limited access, which prevents external validation and, consequently, the scaling up of the findings that have been reported to date [48],[43]. Turning these challenges into non-existing ones is

really important if we want to keep up with the current trend of progress and if we want to keep detection tools that are scientifically credible and ethically sound at the same time.

2.7 SUMMARY AND RESEARCH GAPS

This chapter has outlined the history, the ethical issues, and the technological progress in identifying text produced by AI, particularly in science and academia. It started with the past natural language generation (NLG) and how much more advanced the large language models (LLMs) like GPT, BERT, and RoBERTa are becoming. These advances have significantly changed the process of content development; simultaneously, they have posed complex issues related to the proof of authorship and academic honesty. Although the focus of the early studies was on the linguistic uniqueness of the AI-generated text, the quick enhancement of coherence, context reasoning, and stylistic imitation has made it hard to differentiate between human and machine writing, which has called for more powerful detection systems [2],[3],[1].

An in-depth study of previous works uncovered significant improvements in the methodology throughout the three main phases of research. First, Logistic Regression, SVM, and Random Forest, which are traditional machine learning models, managed to achieve initial success with superficial linguistic and stylometric features but their poor contextual understanding limited the generalization through different datasets [57],[13]. Secondly, the rise of transformer-based architectures mainly BERT, RoBERTa, and SBERT marked the beginning of the period of representing contextual features and semantic modeling which resulted in very high detection accuracy over a wide range of text domains [58],[49],[14]. Thirdly, the researchers of the recent papers started to create hybrid and ensemble learning frameworks, combining statistical and semantic approaches to improve performance, interpretability, and ethical transparency [24],[47],[54].

This chapter further emphasized the increasing significance of XAI and dataset diversity in the construction of reliable detection systems. Research that applied techniques like SHAP and LIME has shown that it is possible to interpret the predictions made by models, but these methods are still not widely applied in the area of academic detection [13],[54]. Similarly, the formation of equal and specific to the domain datasets like AI-GA, HWAI, and HAGT-1M is a very important measure toward the development of reproducible

benchmarks. However, many existing datasets are still concentrating on social media or general text sources rather than academic prose, thus there is a range of scholarly content that is still unassessed regarding evaluation.

Significant breakthroughs have been made in the field, but still, there are some research gaps which are very much in evidence:

- a. **Insufficient Domain Specific Benchmarking:** Nearly all the detection frameworks depend on non-specific datasets instead of pertaining academic corpora which lead to less evaluation of the writing's unique linguistic and rhetorical features [20],[47].
- b. **Shortage of Interpretability and Openness:** Even though the deep learning and transformer-based models achieve great accuracy, their decision-making is very often kept hidden. Very few studies focus on explainability or utilize interpretable hybrid models which are fit for academic use [13],[54],[43].
- c. **Inadequate Application of Linguistic and Statistical Properties:** The majority of the latest detection strategies function with embeddings or n-gram features exclusively, failing to consider such salient linguistic indicators like coherence, lexical variation, or syntactic composition that might enhance detection reliability [7],[59].
- d. **Limited Cross-Domain Testing:** The opportunity for detection models to generalize their application to various research areas and text types is still uncharted. A number of models exhibit domain dependency, showing good outcomes only for data that is very much like their training corpus [22],[60].
- e. **Ethical and Academic implications:** The consequences of the automated authorship detection, such as fairness, possible bias against non-natives, and implications for peer review and academic freedom have been mentioned in very few studies and those few studies have not covered the topic in a very comprehensive way [19],[52].

In order to fill these research gaps, the current research presents a double-scenario framework that merges transformer-based embeddings (Scenario 1) and ensemble learning with linguistic and statistical analysis (Scenario 2) as a whole. This strategy is intended to provide AI-GA, HWAI, and HAGT-1M datasets with interpretability, scalability, and robustness at the same time. The combination of the innovation in methodology and moral responsibility results in the development of the transparent, reproducible, and high-accuracy detection models that not only support but also strengthen the trustworthiness of scholarly communication in the era of artificial intelligence.

3 METHODOLOGY

3.1 INTRODUCTION

The current chapter illuminates the methodological framework which was utilized to reach the aims of this research, which primarily revolves around the discrimination and identification of AI-generated and human-written academic texts. The methodology is constructed in such a way that it guarantees fairness, reproducibility, and interpretability, all of which are necessary in overcoming the technical and ethical issues that accompany AI-produced scholarly texts, as depicted in Figure (3.1).

For this purpose, the research is laid out in two interrelated experimental scenarios. The initial scenario takes advantage of a transformer-based framework that utilizes contextual embeddings from cutting-edge language models like SBERT and RoBERTa, coupled with classic machine learning classifiers. This setup sets a highly effective baseline for revealing the stylistic and contextual differences that exist between human writing and that produced by AI.

The second scenario adds to the first by using the techniques of ensemble learning, linguistic analysis, and exploratory data analysis (EDA) to increase generalization, interpretations, and robustness. This method of detection not only employs several classifiers via ensemble voting as well as boosting but it also provides a way of text authenticity detection that is open and flexible especially for the academic sector.

Both cases are tested under the same experimental conditions with three datasets that were specially selected to give a fair and unbiased comparison. In addition to accuracy, precision, recall, F1-score, and ROC-AUC, the Friedman and Wilcoxon tests are applied as statistical validation methods to confirm the significance of the performance differences.

In general, this method gives a thorough base for the evaluation of the strengths and weaknesses of the two methods, transformer-based and ensemble-based, to going forward with this discussion. Moreover, it offers a way of discussion about academic integrity that is able to be repeated and is ethically sound through the framework for distinguishing between AI-generated and human-generated scholarly content.

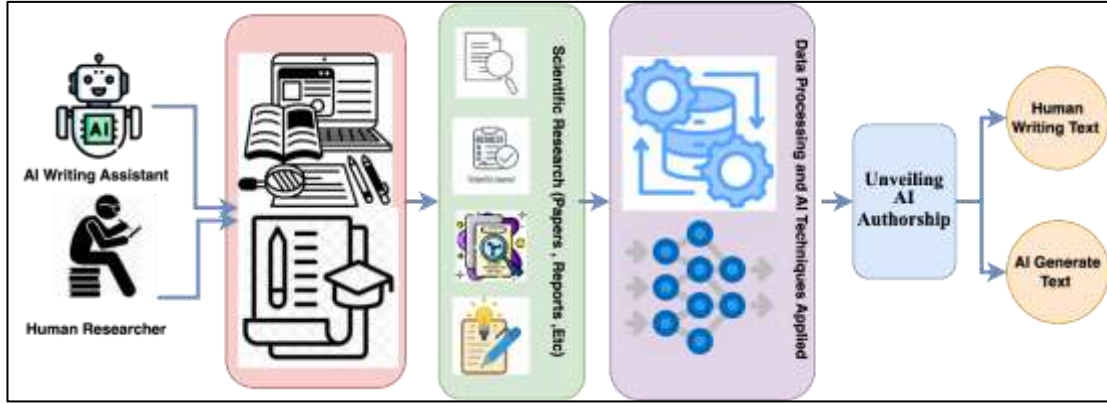


Figure 3.1: Conceptual Framework of the Proposed Methodology for AI-generated Academic Text Detection.

3.2 SOFTWARE USED

This study reports using Anaconda3 (version 2.5.2), which is the open-source Python distribution for data science, as the platform on which all the classification models and their outputs were realized. Anaconda combines a package manager, environment manager, and more than 1,000 pre-installed scientific libraries and frameworks into an integrated environment. It is also cross-platform and can run on Windows, macOS, and Linux, thus being a powerful and great tool for research and academic computing (see Figure 3.2).

Anaconda Navigator provides a user-friendly graphical interface to launch the development tools such as Jupyter Notebook, Spyder, and JupyterLab, which were the main tools in this study. The analyses, model development, and evaluation metrics like accuracy, precision, recall, and F1-score were done using Python libraries that come with Anaconda, such as Scikit-learn, NumPy, Pandas, and Matplotlib [61].

In academic research environments, Anaconda Enterprise offers an array of benefits including:

- a. Version control for data science assets and reproducible environments.
- b. Secure access management and detailed system activity logging.
- c. Integration with institutional authentication systems.
- d. Access to a vetted open-data science ecosystem that supports ethical and transparent research practices.

Anaconda's Python notebooks were the main research environment for experimentation throughout this study because of their flexibility, comprehensive library support, and excellent performance in text mining, feature engineering, and model evaluation.

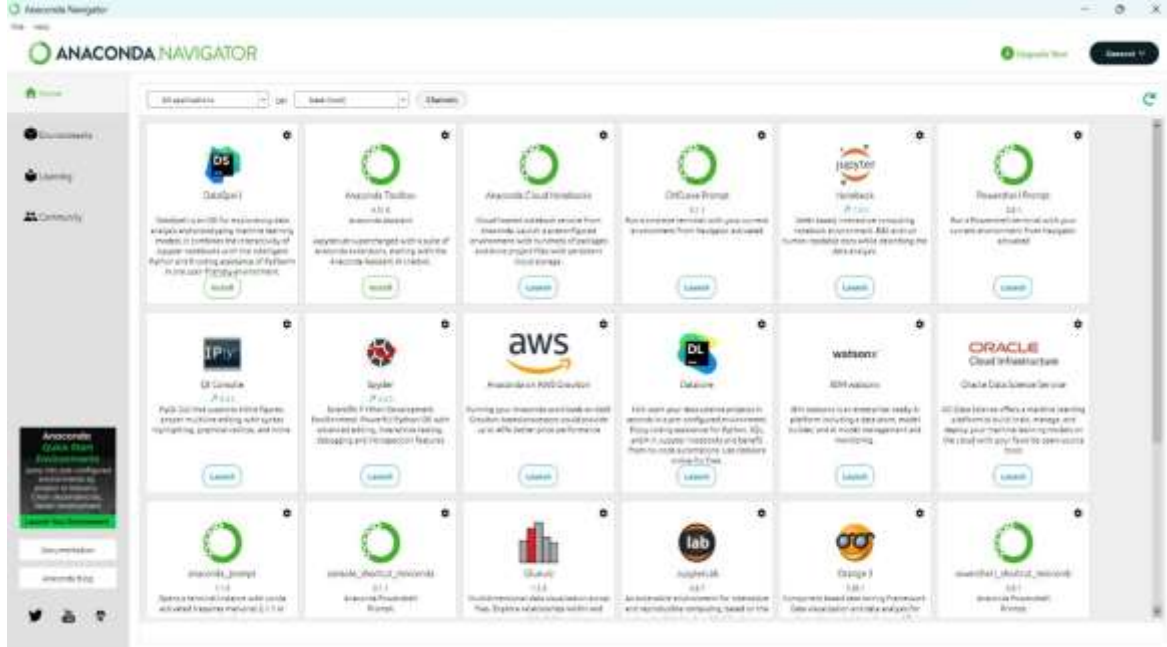


Figure 3.2: Interface of the Anaconda Navigator Environment Used in This Study.

3.3 DATASETS DESCRIPTION

The research in question uses three datasets that complement each other to showcase different styles of academic writing and AI-generated content from different disciplines as well. The datasets were selected carefully in such a way that they would all be equally distributed, rich in vocabularies, and typical of the real scholarly writing. Each dataset has its own strength in terms of testing the reliability and applicability of the proposed detection frameworks in both experimental situations.

3.3.1 AI-GA Dataset (Academic Abstracts)

The AI-GA dataset was derived from the free-to-use COVID-19 Open Research Dataset (CORD-19), an extensive collection of peer-reviewed and preprint academic papers on COVID-19 and other coronaviruses. CORD-19 was launched to support text mining and natural language processing (NLP) research within scientific communication and knowledge discovery [52].

A subset of 14,331 English-language papers with both the title and the abstract was picked up randomly from the corpus. Each abstract was matched with an AI-generated version that was produced using OpenAI’s GPT-3.5 (Davinci) model.

To make the text diverse and original, temperature (randomness) and frequency penalties were varying, while prompt words and repeated tokens were excluded. At the end, a well-distributed open-access collection of 28,662 entries was created, which consists of 14,331 human-written and 14,331 AI-generated abstracts.

Each entry in the dataset has three fields: title, abstract, and label, as shown in Figure 3.3. The label attribute is binary encoded, with 0 for human-written and 1 for AI-generated texts.

	title	abstract	label
0	Clinical features of culture-proven Mycoplasma...	OBJECTIVE: This retrospective chart review des...	0
1	Nitric oxide: a pro-inflammatory mediator in L...	Inflammatory diseases of the respiratory tract...	0
2	Surfactant protein-D and pulmonary host defense	Surfactant protein-D (SP-D) participates in th...	0
3	Role of endothelin-1 in lung disease	Endothelin-1 (ET-1) is a 21 amino acid peptide...	0
4	Gene expression in epithelial cells in respons...	Respiratory syncytial virus (RSV) and pneumoni...	0
...	—	...	—
28657	Characteristics of the Dynamic of Mobile Networks	\n\nThis empirical study examines the characte...	1
28658	Prediction and Prevention: Interventions to En...	\n\nThis journal articulates the impact of pro...	1
28659	Asthma and Microbes: A New Paradigm	\n\nAsthma is a chronic inflammatory disease o...	1
28660	Diagnostic Knowledge in the Genetic Economy an...	\n\n\nThis paper proposes to investigate the ...	1
28661	High Capacity Vectors	\n\nThis study entitled 'High Capacity Vector...	1

Figure 3.3: Snapshot of the AI-GA Dataset Structure Showing the Fields: Title, Abstract, and Label.

A multi-stage preprocessing pipeline was applied to ensure text quality and consistency, including:

- Removal of special characters, HTML tags, and non-alphabetic symbols.
- Elimination of redundant whitespace and line breaks.
- Lowercasing of all text to maintain uniformity.
- Stop-word filtering to remove non-informative terms (e.g., the, and, of).
- Most common generic terms (i.e. paper and study) were excluded because of their extreme representation in AI texts.

The balanced class distribution of the dataset is illustrated in Figure 3.4, which corroborates the equal representation of the AI and human abstracts. The balance provides the opportunity for model training and evaluation to be fair, thus minimizing the bias in classification.

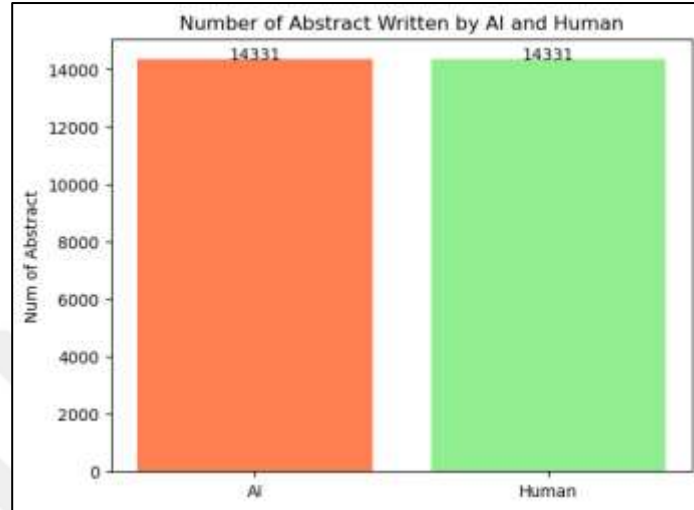


Figure 3.4: Distribution of AI-generated and Human-written Abstracts in the AI-GA Dataset.

In general, the AI-GA dataset functions as a domain-specific standard for the scrutiny of linguistic, structural, and stylistic alterations between genuine academic abstracts and those produced by AI. It offers a substantial basis for the assessment of text authenticity models in scholarly settings.

3.3.2 HWAI Dataset (Human–AI Academic Writing Corpus)

The HWAI dataset consists of 44,162 articles evenly distributed between human-generated (HGT) and machine-generated (MGT) classes, thus offering a solid base for testing AI authorship detection models. Each entry contains two main fields Text and Class, depicted in Figure 3.5. The Class column is coded as 0 for human text and 1 for machine-generated text [48].

	Text	Class
0	Mississippi Highway 403 and MS 15 in Mathiston...	0
1	State Route 97 northwest of Congress northeast...	0
2	Edwin B. Erickson III was an American politici...	0
3	The Tiger Fire was a wildfire that burned 16,2...	0
4	Fajsz , was Grand Prince of the Hungarians fro...	0
...	...	...
9995	Women in early modern Scotland, between the Re...	0
9996	The Court of Common Pleas was one of the four ...	1
9997	Megara is a fictional character from the Disne...	1
9998	Luan Da is a Chinese literary figure who lived...	1
9999	The New York Herald Tribune was a newspaper pu...	0

44138 rows x 2 columns

Figure 3.5: Sample Structure of the HWAI Dataset Showing the Text and Class Fields Representing Human and AI-generated Articles.

The visual representation of the dataset's structure, which is shown in Figure 3.5, not only emphasizes a well-organized system but also supports the implementation of supervised classification tasks. It exhibits the various kinds of texts and themes that the Wikipedia-based corpus covered, such as history, science, and general knowledge subjects.

The visual representation of the distribution as shown in Figure 3.6 verifies the balanced nature of the dataset with 22,081 human-written articles and 22,069 AI-generated articles. This almost perfect balance skews no influence during training of the model and at the same time, ensures reliable performance evaluation over all experimental scenarios.

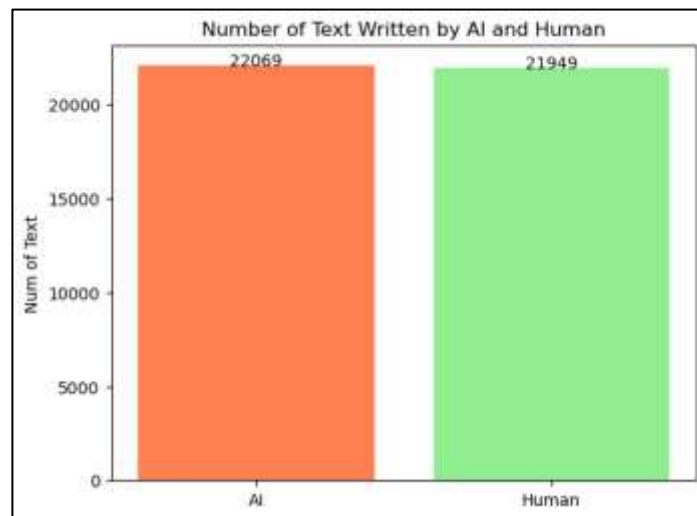


Figure 3.6: Distribution of Human-written and AI-generated Articles in the HWAI Dataset.

3.3.3 HAGT-1M Dataset (Human–AI Generated Text Corpus)

The HAGT-1M dataset is a large-scale benchmark designed for the discrimination between human and AI-generated academic text. This dataset is composed of around one million text samples, with a perfect split between the two categories of 500,000 each, this guarantees a perfectly balanced structure for large-scale model training and evaluation. Additionally, the dataset adds up to the AI-GA and HWAI corpora by its linguistic variety and encompassing various disciplines' general academic writing styles [24]. For this study, a balanced subset of 500,000 samples was picked, which included 250,000 human and 250,000 AI-made texts. Besides, the selection guarantees that the computational requirements remain manageable while the data remains representative and balanced.

HAGT-1M's human-written part was formed from the texts that are freely available in the online academic repositories such as arXiv, ScienceDirect, and IEEE Xplore, that are covering the areas of computer science, information technology, and applied sciences. The filtering of articles was done in a very meticulous manner so that only the abstracts and the research sections of the English-language were included based on their quality and length standards, thereby guaranteeing uniformity of both textual complexity and vocabulary density.

Each record in the HAGT-1M dataset consists of four primary fields:

- a. ID: A unique identifier assigned to each sample.
- b. Human_Text: The original text written by a human author.
- c. AI_Text: The corresponding text generated by an AI model.
- d. Instructions: The context or prompt used to generate the AI text.

The human-written content was compiled from open academic repositories and online writing resources, including educational essays, short reports, and scientific abstracts, while the AI-generated texts were created using OpenAI's GPT-based models, including GPT-3 and GPT-3.5.

A small portion of the data sample is depicted in Figure 3.7, which shows the dataset's truncation and its structure consisting of text and class labels. The label column applies a

binary classification method, in which 0 signifies human-written text and 1 indicates AI-generated text.

	text	label
0	Even though lots of students like to work at h...	0
1	Fresh air we're use to being home all day and ...	0
2	Many people believe that self-esteem comes fro...	0
3	Clearly, everyone should always aim high and a...	0
4	I really think that can be really stressed for...	0

Figure 3.7: Sample Structure of the HAGT-1M Dataset Showing Text Entries and Binary Labels.

Figures 3.7 and 3.8 depict the structure and the distribution of the dataset. In the preprocessing stage, the sample data structure is represented by Figure 3.7, whereas the number of AI-generated and human-written texts is shown in Figure 3.8. The visual corroborates that the classes are almost perfectly balanced, which is a necessary condition for avoiding bias during model training and evaluation.

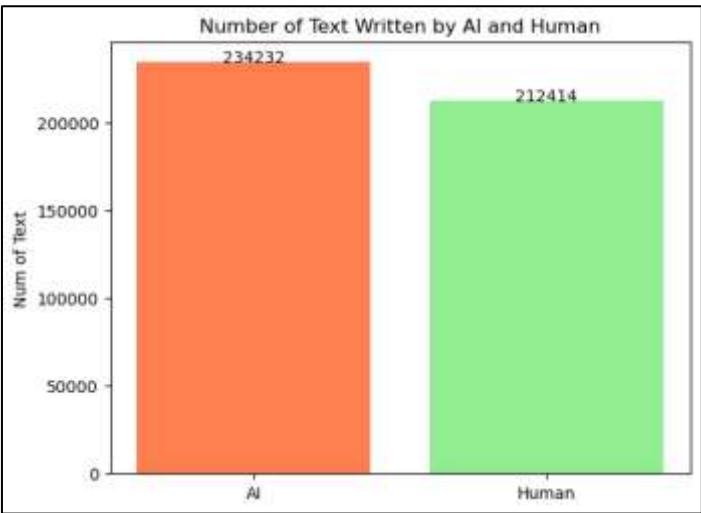


Figure 3.8: Distribution of AI-generated and Human-written Text Samples in the HAGT-1M Dataset.

This research has relied on three datasets: AI-GA, HWAI, and HAGT-1M. Together, they build a strong basis for assessing AI-generated text detection in the academic and linguistic dimensions. The AI-GA dataset gives the input of specific domain scholarly abstracts with a focus on the scientific writing patterns, and the HWAI dataset, on the other hand,

contains encyclopedic and general academic narratives with a range of word-length structures. The HAGT-1M dataset, in contrast, offers a vast and diverse collection of text samples that facilitate generalization and model evaluation's robustness. These datasets together present different angles that mirror both the specialized and generalized academic writing styles, thus providing a fair and comprehensive testing of the transformer-based and ensemble-based frameworks. The use of such a multi-dataset approach guarantees that the resulting detection models are not only precise but also suitable for a broad range of real-world academic content.

3.4 SCENARIO 1: TRANSFORMER-BASED DETECTION FRAMEWORK

3.4.1 Overview

The proposed case embodies the fundamental experimental structure for distinguishing between human and AI-generated academic texts through the combination of transformer-based embeddings and classic machine learning classifiers. The intention behind this configuration is to create a solid reference point for measuring model performance, interpretability, and scalability before moving to ensemble-based techniques, as illustrated in Figure 3.9.

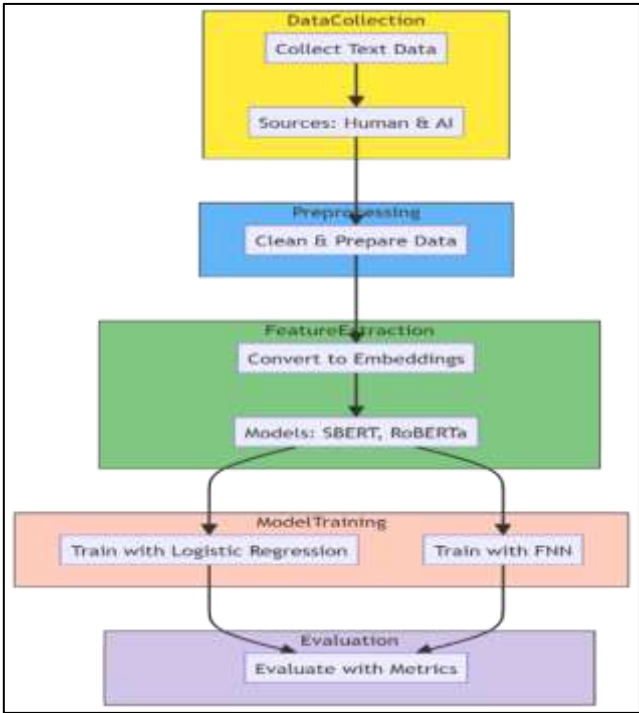


Figure 3.9: Framework of the Transformer-Based Detection Model.

The framework takes advantage of Sentence-BERT (SBERT) and RoBERTa to transform text into two, which are both the most prominent architectures for understanding the context of the text to produce high-dimensional vectors that are able to depict the semantic and syntactic subtleties present in the human and AI-generated academic content. The vectors produced are an extremely compact and rich representation of the sentences, which allow for powerful classification in the following steps.

Afterwards the embeddings were extracted, two classifiers Logistic Regression (LR) and a Feedforward Neural Network (FNN) were trained on the supervised approach to classify the text into two categories. Logistic Regression was deemed the best candidate for its transparency and the property of the embeddings being linearly separable, while the FNN was chosen to deal with the more complicated, non-linear relationships.

The model went through the process of training and validation using three different datasets (AI-GA, HWAI, and HAGT-1M) under the same preprocessing and partitioning conditions to present fair experiments. Every dataset was split into two parts, where 80% was used for training and 20% for testing. In order to give a thorough overview of the classification reliability, the model was evaluated by means of accuracy, precision, recall, F1-score, and ROC-AUC metrics.

The entire operation of the transformer-based framework is depicted in Algorithm 3.1, which illustrates the sequential activities from data loading and embedding generation to classification and evaluation.

Algorithm 3.1: Transformer-Based Detection Procedure.**Input:** Preprocessed academic text corpus.**Output:** Classified text labels (AI-generated or human-written).**Begin****Step 1: Data Input**

Load preprocessed text datasets (AI-GA, HWAI, HAGT-1M).

Step 2: Embedding GenerationFor each text sample t in the dataset:Tokenize t using SBERT and RoBERTa tokenizer.Obtain contextual embeddings $E_SBERT(t)$ and $E_RoBERTa(t)$.Concatenate embeddings to form $E_total = [E_SBERT \oplus E_RoBERTa]$.**Step 3: Model Training**

Split dataset into 80% training and 20% testing subsets.

Train Logistic Regression (LR) and Feedforward Neural Network (FNN) classifiers on E_total .**Step 4: Model Evaluation**

Predict on test data and compute performance metrics:

Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Compare LR and FNN performance to identify baseline results.

End

In general, this particular case points out the ability of transformer-based text representations to distinguish between AI and human authorship in scholarly writing. It also plays the role of providing both the conceptual and experimental support for the second case, which makes the use of ensemble learning to increase interpretability and generalization in the various academic fields the main focus.

3.4.2 Text Representation

In AI detection of academic writing, text representation plays a very important role as it acts to determine the means through which textual information is encoded numerically and interpreted by algorithms. In the present research, one of the methods of semiconductor representation is through the use of Sentence-BERT (SBERT) and RoBERTa, which are two transformer-based architectures that are expert-known for their capability to yield context-aware embeddings that encompass both syntactic and semantic dimensions.

SBERT is a variant of the BERT network that is specifically engineered for generating semantically meaningful sentence embeddings. It employs a Siamese network structure to

derive sentence-level representations that can be directly compared using cosine similarity or fed into downstream classifiers [56]. RoBERTa, a top-performing BERT variant, enhances training robustness by getting rid of the next sentence prediction objective and dynamically varying the masking patterns, thus yielding higher-quality contextual embeddings [58]. The simultaneous use of these models guarantees that there will be a comprehensive and nuanced understanding of the linguistic patterns present in the academic texts.

The text from the three datasets AI-GA, HWAI, and HAGT-1M was processed in the same manner by the same model, and therefore their tokenization was done likewise. The sequences of tokens formed through tokenization were later provided to the pretrained SBERT and RoBERTa models in order to get the 768-dimensional embedding vectors [28]. During this process, the embeddings were fused to create a joint representation vector (E_{total}), which incorporated both global and local context dependencies.

The embedding generation process can be formally written as:

$$E_{total} = [E_{SBERT} \oplus E_{RoBERTa}] \quad (3.1)$$

Where E_{SBERT} and $E_{RoBERTa}$ refer to the sentence embeddings received from the two transformer models, and $\oplus$ denotes the concatenation of vectors. This merged embedding matrix acts as the input for the succeeding classifiers Logistic Regression and Feedforward Neural Network, as illustrated in Figure 3.10, that are used in the next step.

In fact, the framework's dual-embedding strategy which the authors came up with introduced the different AI text and human-written academic prose [25]. as the very fine details of linguistic variations thus paving the way for the better-insight and around the clock oversight of AI text and human-written academic prose. Moreover, the technique makes it possible for the model to be of general application over the datasets that differ in domains, writing styles, and text lengths.

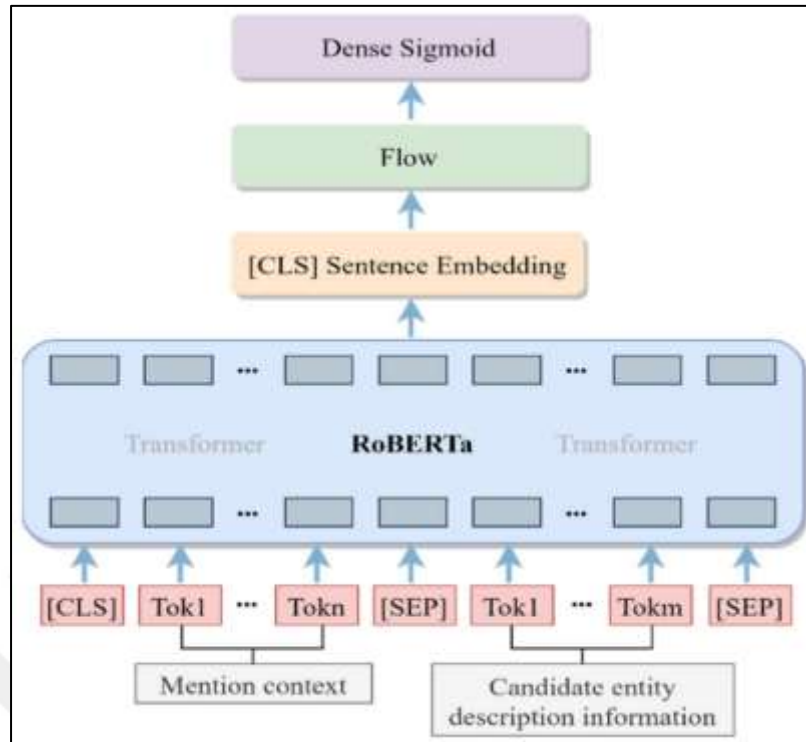


Figure 3.10: Architecture of SBERT and RoBERTa for Generating Sentence Embeddings.

The architectural design of the Sentence-BERT (SBERT) and RoBERTa models that were chosen for text embedding generation during this research is shown in Figure 3.11. Both models are transformer encoder-based but differ in the ways they are trained and their optimization approaches. RoBERTa, for example, enhances BERT’s masked language modeling by constantly changing the masking patterns and eliminating the next-sentence prediction task, which leads to richer contextual embeddings. SBERT, on the other hand, not only fine-tuning BERT to generate semantically meaningful sentence-level embeddings through a pooling operation over token embeddings but also enables efficient comparison of sentence similarity. These embeddings are then used as input features for the classification models that will eventually distinguish the AI-generated academic texts from the human-written ones.

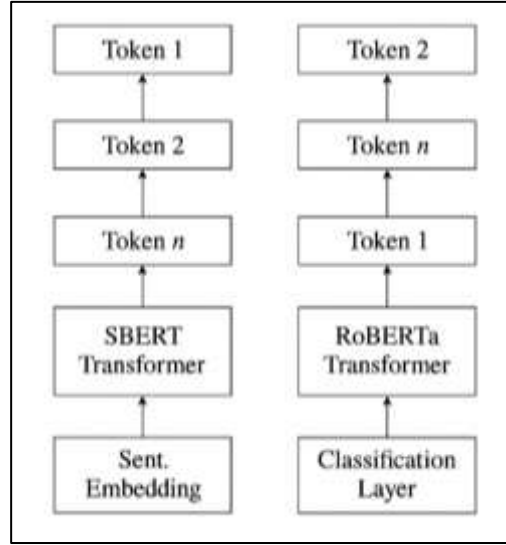


Figure 3.11: Architecture of SBERT and RoBERTa.

3.4.3 Classification Models

Once the embeddings were created through Sentence-BERT (SBERT) and RoBERTa, two supervised classification methods were used to differentiate the AI-generated and human-written academic texts: Logistic Regression (LR) and Feedforward Neural Network (FNN). Choosing these models yielded a fair comparison between a traditional linear classifier and a neural one that could recognize non-linear ties in the embedding space.

3.4.3.1 Logistic Regression (LR)

Due to its simplicity, interpretability, and robustness in handling high-dimensional vector spaces like sentence embeddings, Logistic Regression is a strong baseline for text classification among the other classification methods [62]. The classifier provides the probability that the given text is of a certain class (AI-generated or human-written) through the sigmoid activation function:

$$P(y = 1 | x) = \sigma(w^T x + b) \quad (3.2)$$

here x is the embedding vector, w and b are the parameters of the model learned during training, and $P(y = 1 | x)$ is the probability predicted for the text to be AI-generated. The procedure of optimization of model parameters is done by means of binary cross-entropy loss which guarantees convergence towards the maximum likelihood estimate.

The optimal regularization parameter C , which is the one in between underfitting and overfitting, was found through grid search with cross-validation. This method resulted in a steady performance for all three datasets namely AI-GA, HWAI, and HAGT-1M.

3.4.3.2 Feedforward Neural Network (FNN)

Their goal was to employ a Feedforward Neural Network to identify and represent the intricate, non-linear boundaries of decision-making situated in the embedding space [63]. The setup includes:

- a. An input layer corresponding to the embedding dimension (e.g., 768 for SBERT and RoBERTa).
- b. One or two hidden layers with ReLU activation functions to introduce non-linearity.
- c. A sigmoid output layer for binary classification.

The model was trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 32. Early stopping was implemented to prevent overfitting, monitoring validation loss with a patience of 5 epochs.

Formally, the forward propagation process in the FNN can be represented as:

$$h_1 = f(W_1x + b_1), h_2 = f(W_2h_1 + b_2), \hat{y} = \sigma(W_3h_2 + b_3) \quad (3.3)$$

where $f(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ is the sigmoid activation applied to the final output layer.

3.4.3.3 Training configuration

The training of both classifiers was done in the same experimental setup:

- a. Training/Test Split: 80% / 20%.
- b. Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
- c. Validation Strategy: 10-fold cross-validation to guarantee statistical reliability.
- d. Computational Environment: The whole process was done using Anaconda3 (Python 3.10) on a computer with Intel i7 CPU, 16 GB RAM, and NVIDIA RTX GPU acceleration.

3.4.3.4 Summary

The combined application of Logistic Regression and FNN gives the user both the benefit of being able to understand the model as well as the advantage of using very powerful and complex models. The LR classifier presents clear and probabilistic boundaries, whereas the FNN allows the identification of subtle and complex dependencies among the text features. The models in tandem constitute the primary classification phase in the transformer-based detection setup, as depicted in Figure 3.11 and procedurally described in Algorithm 3.1.

3.4.4 Evaluation and Performance Metrics

For the purpose of guaranteeing an equitable and exhaustive evaluation, it was decided to assess both the transformer-based models, i.e. Logistic Regression (LR) and Feedforward Neural Network (FNN), by means of the same metrics and validation procedures. The evaluation framework is centered on evaluating classification performance, stability, and reliability over the three datasets (AI-GA, HWAI, and HAGT-1M).

3.4.4.1 Performance metrics

In order to standardize the metrics mentioned above, the following metrics of performance of the model were used:

Accuracy: This is a measure of how many of the total samples were classified correctly.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.4)$$

Here, TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

Precision: This metric reflects the performance of the model in predicting the positive class.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3.5)$$

Recall: The model's capability to accurately flag the positive ones is expressed through this metric.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3.6)$$

F1-Score: Represents the harmonic mean of Precision and Recall, providing a balanced measure between them.

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.7)$$

ROC-AUC: (Receiver Operating Characteristic and Area Under the Curve) This metric reveals how well the model can separate texts written by humans from those generated by AI, regardless of where the classification threshold is set. The higher the AUC, the better the separation [64].

$$AUC = \frac{U_1}{n_1 n_2} \quad (3.8)$$

Where (U_1) indicates the (U) value that was calculated using sample 1, (n_1) indicates the number of sample 1, and (n_2) is the number of sample 2 (sample 2 is the set of examples that were not chosen to be in sample 1).

3.4.4.2 Validation strategy

To guarantee that the model was durable and could be generalized, each model was tested with 10-fold cross-validation [65]. This means that the dataset was split into 10 equal parts; out of these, 9 were used for training, and 1 for testing, thus doing it iteratively until each fold was utilized as a test set once. This tactic results in reducing the error of estimate due to data partitioning, and also in getting a more reliable estimation of the model performance.

The steps of the evaluation were:

- a. Each dataset was divided into two parts: 80% for training and 20% for testing.
- b. 10-fold cross-validation was applied to the training set.
- c. Individual metrics were recorded for each fold for accuracy, precision, recall, F1-score, and ROC-AUC.
- d. Mean and standard deviation were computed to check the model's stability and consistency.

3.4.4.3 Computational environment

All the experiments were carried out in Python 3.10 using Anaconda Navigator (Version 2.5.2) on a machine with the following specifications:

- a. Processor: Intel® Core™ i7, 3.4 GHz
- b. RAM: 16 GB DDR4
- c. GPU: NVIDIA GeForce RTX 3060 (6 GB VRAM)
- d. Operating System: Windows 11 (64-bit)

The principal Python libraries that were used are NumPy, Pandas, Scikit-learn, TensorFlow, and Matplotlib. To maintain uniformity, every experiment was performed thrice, and the random seeds were set for the sake of reproducibility.

3.4.4.4 Summary

With its strong points and weak points interpretation, this evaluation framework gives a fair and clear view of performance of models like Logistic Regression (interpretable) and FNN (complex). The application of several datasets, separation done in the same way each time, and use of statistics for evaluation all together make not only the results to be reliable but also distributable across different and even opposite academic text domains to a certain extent.

3.5 SCENARIO 2: ENSEMBLE LEARNING-BASED DETECTION FRAMEWORK

3.5.1 Overview

The advanced stage of the research framework that has been developed to improve interpretability, robustness, and generalization in the detection of AI-generated academic texts is portrayed here. This scenario uses the transformer-based method from the previous scenario as a starting point and then combines the use of ensemble learning, linguistic analysis, and exploratory data analysis (EDA) to come up with the new detection method that is more transparent and adaptable to different academic disciplines.

The deep transformer embeddings typically represent black-box models, while the ensemble learning approach, on the other hand, focuses on feature-level explain ability.

Moreover, it employs TF-IDF vectorization and linguistic indicators such as word frequency, n-gram distribution, sentence length, and syntactic patterns in order to identify the fine-grained stylistic and structural differences that exist between human and AI-generated writing. These interpretable features not only provide but also indirectly lead to the recognition of AI-generated academic content through its vocabulary richness, cohesion, and contextual flow. Figure 3.12 displays the overall architecture of the framework that consists of major stages of data preprocessing, feature extraction, classifier training, and ensemble fusion [23].

The ensemble architecture employs a Voting Classifier mechanism to bring together a variety of supervised classifiers, namely Logistic Regression (LR), Support Vector Machine (SVM), Extremely Randomised Trees (ErT), Random Forest (RF), Stochastic Gradient Descent (SGD), XGBoost, and AdaBoost. The ensemble, by pooling together the predictive strengths of the various models, attains a significantly higher level of stability and is less prone to the biases of the dataset. Thus, the hybrid setup guarantees the same level of performance even on different academic datasets that are of varied nature.

Moreover, the framework makes use of the Friedman test and Wilcoxon signed-rank test as the statistical validation techniques to ascertain the statistical significance of the observed improvements. The tests confirm that the enhanced performance of the ensemble is not a result of random variation but rather a consequence of innovation in methodology.

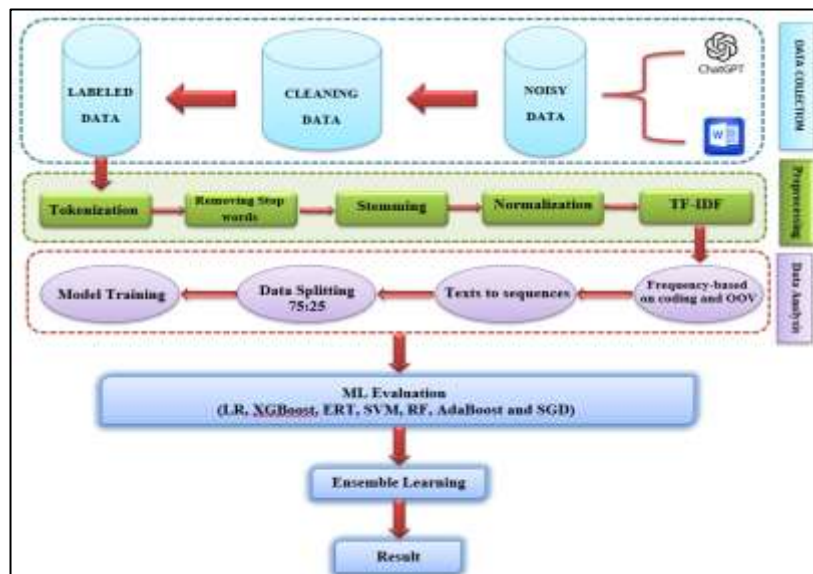


Figure 3.12: Ensemble Learning-Based Detection Framework.

Algorithm 3.2: Unified Framework for AI-Generated Text Detection.

Input: Raw textual datasets (AI-GA, HWAI, and HAGT-1M)

Output: Predicted class label — AI-generated or Human-written text

Begin

Step 1: Data Preprocessing

1. Import raw datasets containing human and AI-generated academic texts.
2. Remove special characters, HTML tags, punctuation, and extra whitespace.
3. Convert all text to lowercase for normalization.
4. Eliminate stop words and overly frequent non-informative words (e.g., “paper”, “study”).
5. Retain only alphabetic tokens to prepare clean input for feature extraction.

Step 2: Feature Extraction

6. Initialize **TF-IDF Vectorizer** with parameters:
 - max_features = 10,000
 - ngram_range = (1,2)
7. Compute term frequency (TF) for each token t in document d :
 $TF(t, d) = \text{Frequency of } t \text{ in } d$
8. Compute inverse document frequency (IDF):
 $IDF(t) = \log(N / DF(t))$
where N is total number of documents, $DF(t)$ is number of documents containing t .
9. Derive **TF-IDF** representation:
 $TF-IDF(t, d) = TF(t, d) \times IDF(t)$
10. Extract linguistic features such as:
 - Average sentence length
 - Word frequency and lexical diversity
 - Syntactic and structural complexity
11. Merge TF-IDF and linguistic vectors into a single feature matrix.

Step 3: Model Training and Classification

12. Split datasets into **training (80%)** and **testing (20%)** subsets.
13. Train individual classifiers:
 - Logistic Regression (LR)
 - Support Vector Machine (SVM)
 - Random Forest (RF), Extremely randomised trees (ErT)
 - Stochastic Gradient Descent (SGD)
 - XGBoost
 - AdaBoost
14. Collect predictions from all classifiers.
15. Build the **Voting Classifier Ensemble**:
 - If *Hard Voting*: select majority class label.
 - If *Soft Voting*: compute average predicted probabilities across models.

Step 4: Evaluation and Validation

16. Compute performance metrics — Accuracy, Precision, Recall, F1-score, and ROC-AUC.
17. Perform non-parametric statistical validation using **Friedman** and **Wilcoxon** tests.
18. Output ensemble model performance results.

End

To sum up, Scenario 2 offers a thorough, understandable, and statistically supported method for identifying AI-produced text in academic contexts. It connects the top-

performing but difficult-to-understand transformer models with the transparent, hardly deployable ensemble systems, which respect fairness, reproducibility, and integrity in academic authorship detection.

3.5.2 Linguistic and Statistical Feature Engineering

The process of linguistic and statistical feature engineering is detailed in this section as it was applied to the three academic datasets (AI-GA, HWAI, and HAGT-1M). By methodically turning the text into numerical features that the machine learning can understand, the aim was to show the differences in style and structure between the two types of text, namely AI and human. The analysis took place in three main steps: (1) Exploratory Data Analysis (EDA), (2) Linguistic Analysis (N-grams and Frequent Words), and (3) Statistical Feature Extraction (TF-IDF).

3.5.2.1 Exploratory Data Analysis (EDA)

Through the Exploratory Data Analysis, a first qualitative estimate of the vocabulary distribution and thematic emphasis across the datasets was done. The word clouds were used as a tool for visualizing the most common words, hence giving a clear view of the vocabulary patterns that are characteristic of human and AI-generated texts [66].

In the case of the AI-GA dataset, the combined word cloud for all abstracts is presented in Figure 3.13. The most frequent terms in the corpus are study, result, and data which are common and reflect the academic nature of the corpus. After that, Figures 3.14 and 3.15 display the human-authored and AI-generated subsets, respectively. The human texts exhibit richer domain-specific vocabulary while the AI-generated texts recur and generalize in terms of expressive qualities [67].

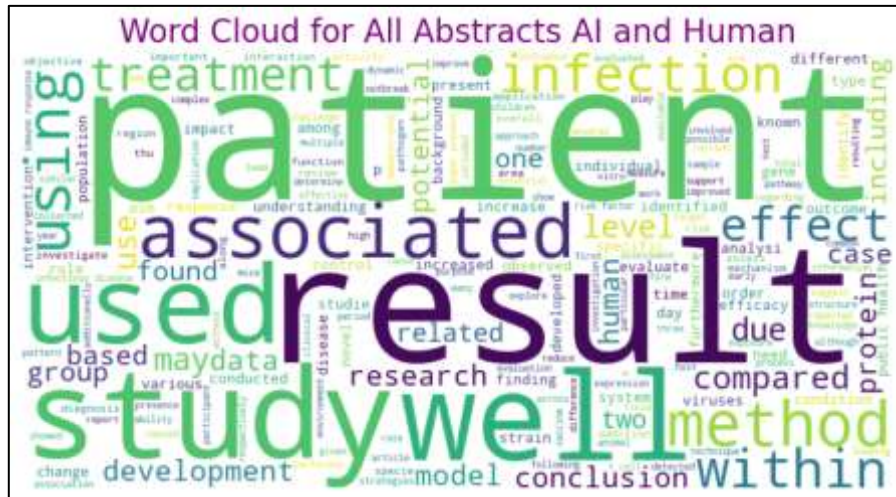


Figure 3.13: Word Clouds for the AI-GA Dataset Combined.



Figure 3.14: Word Clouds for the AI-GA Dataset Human.

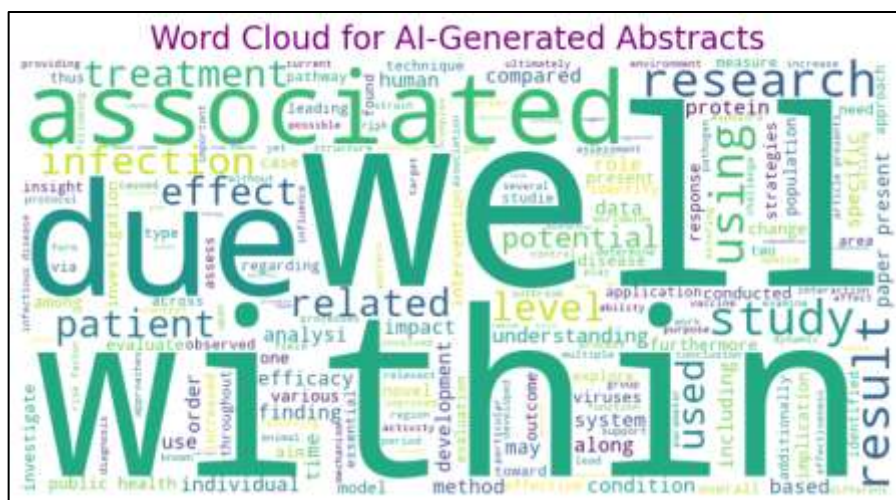


Figure 3.15: Word Clouds for the AI-GA Dataset AI.



Figure 3.18: Word Clouds for the HWAI Dataset AI.

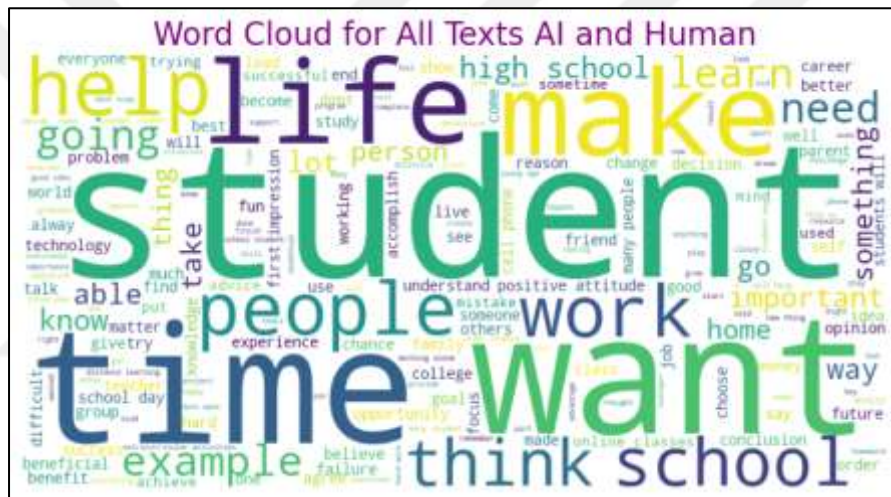


Figure 3.19: Word Clouds for the HAGT-1M Dataset Combined.

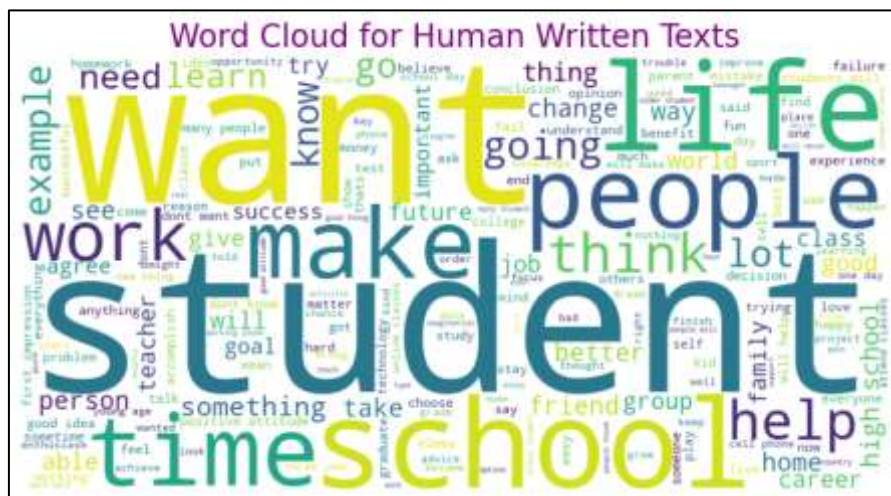


Figure 3.20: Word Clouds for the HAGT-1M Dataset Human.

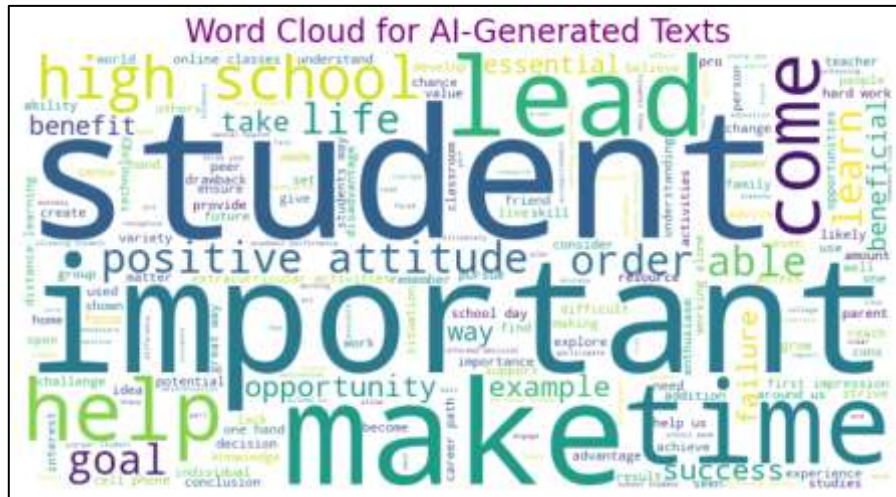


Figure 3.21: Word Clouds for the HAGT-1M Dataset AI.

The results of the investigation showed the same pattern of lexical simplification in different AI texts, which corroborated the assumption that the AI models usually give preference to fluency rather than diversity. In this way, EDA enabled the transition to a more thorough quantitative linguistic analysis in the subsequent stage.

3.5.2.2 Linguistic Analysis (N-grams and Frequent Words)

The structure and variety of expressions are measured by linguistic analysis to a quantitative level, which is done through n-gram frequency distributions, particularly, unigrams (1-grams), trigrams (3-grams), and five-grams (5-grams). This kind of analysis brings out the stylistic inclinations as well as the repetition patterns, which in turn, uncover the characteristic linguistic signatures of human and AI authorship [60].

The illustrations in Figures 3.22–3.24 for the AI-GA dataset point out the differences in linguistic behavior very clearly. The authors from the human side are the ones who usually go for the semantically rich and contextually cohesive phrases, whereas the texts produced by AI are the ones that show the lexical redundancy through the use of connective structures like "presents the results of a" or "this article discusses a." Such continuous occurrences are indeed the signature of the generative model's reliance on structural templates rather than on adaptive academic reasoning.

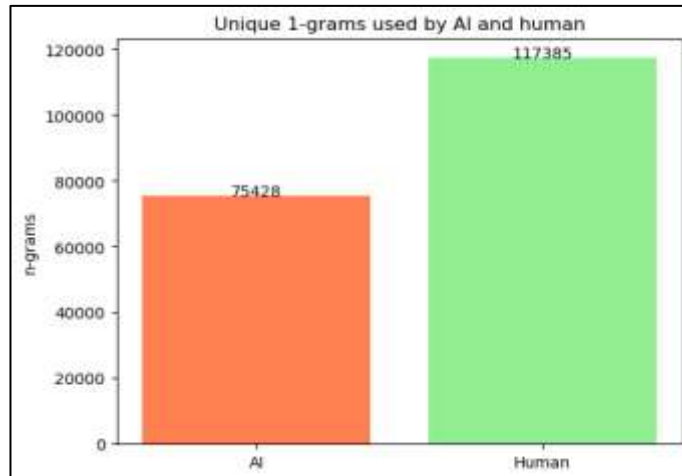


Figure 3.22: Frequency of 1-grams in AI-GA Dataset.

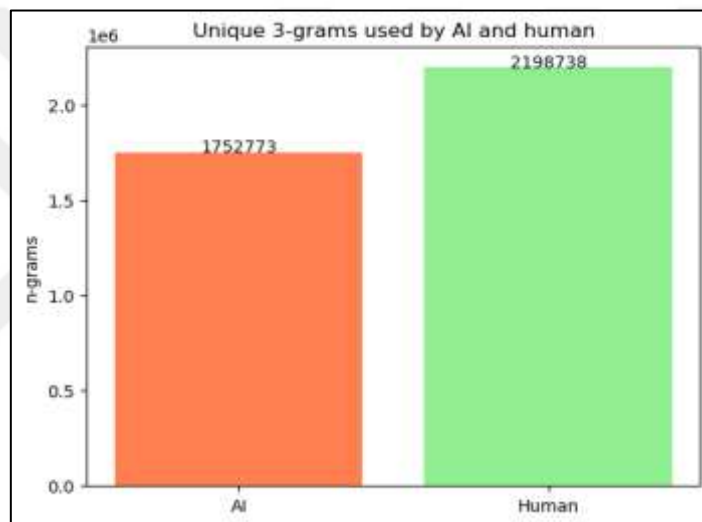


Figure 3.23: Frequency of 3-grams in AI-GA Dataset.

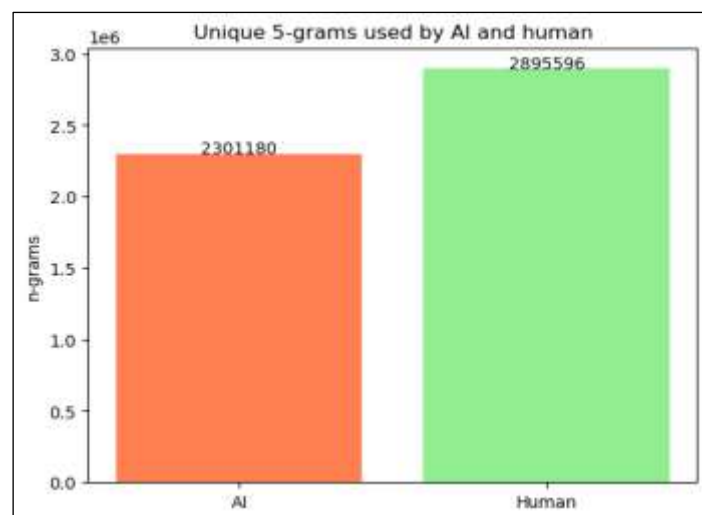


Figure 3.24: Frequency of 5-grams in AI-GA Dataset.

The HWAI dataset (Figures 3.25–3.27) showcases a similar pattern, albeit more pronounced. The human writing has a wider range of vocabulary and employs terminology according to the subject matter, while the AI texts consist mainly of repetitions of general expressions (e.g., “as a reminder of the importance of”). This implies that AI writing in encyclopedia style is dominated by the ease of syntax rather than the variety of vocabulary; this is a situation that is made worse by the practice of open-domain language modeling [55].

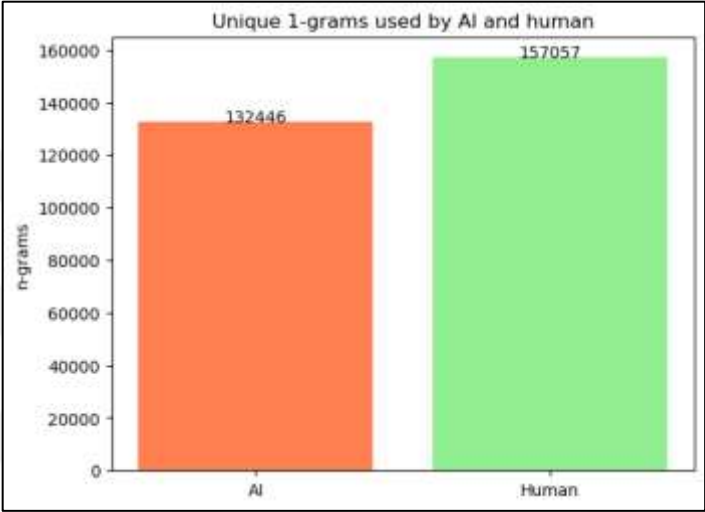


Figure 3.25: Frequency of 1-grams in HWAI Dataset.

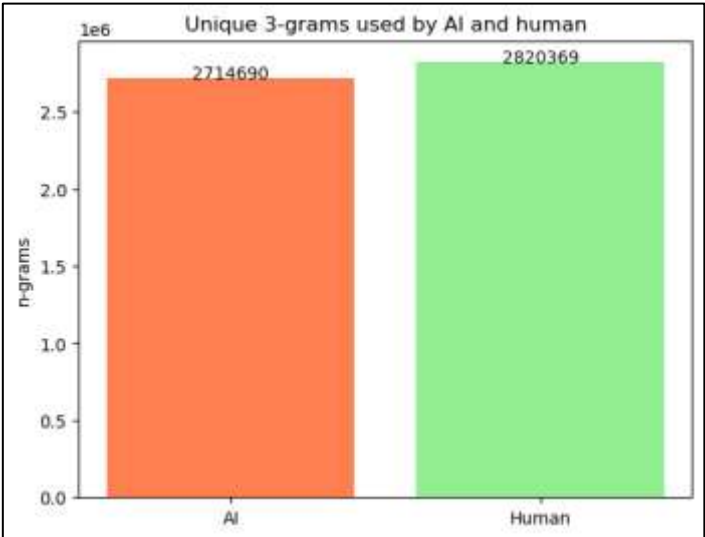


Figure 3.26: Frequency of 3-grams in HWAI Dataset.

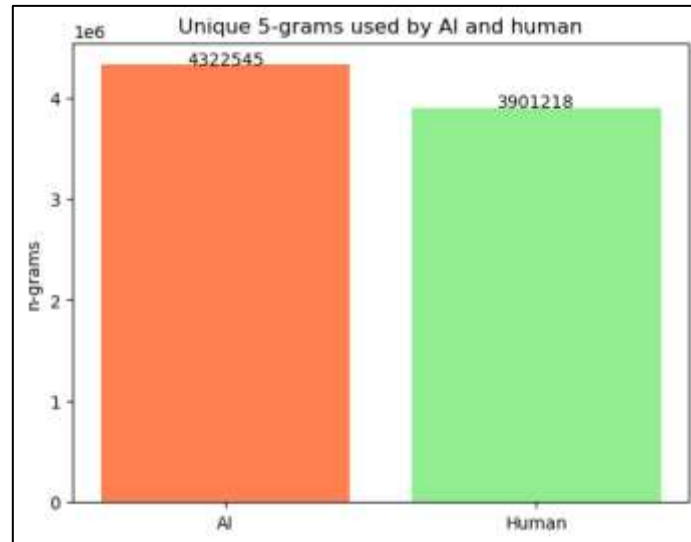


Figure 3.27: Frequency of 5-grams in HWA1 Dataset.

The tendency continues to be the same for the HAGT-1M dataset (Figures 3.28–3.30) which contains academic as well as general argumentative compositions. The AI-generated texts, notwithstanding their grammatical uniformity, are nevertheless characterized by the absence of the human-authored texts' creativity at the phrase-level and diversity in meaning. This reveals that the transformer-based generation, although being very good at coherence, in most cases gives up the novelty in terms of vocabulary [47].

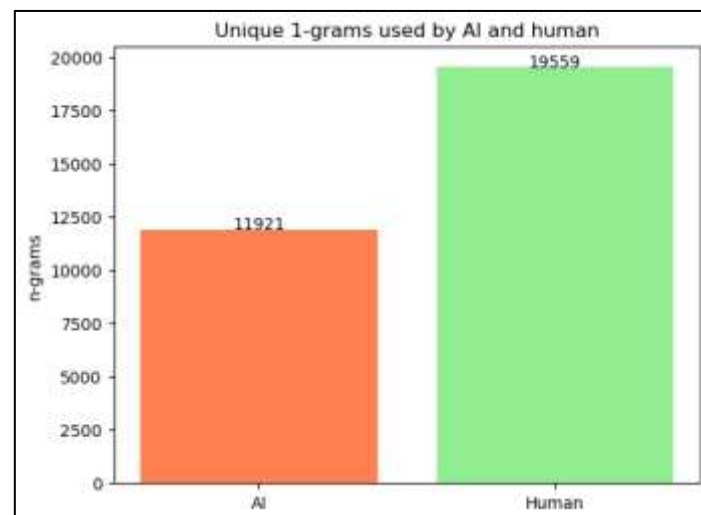


Figure 3.28: Frequency of 1-grams in HAGT-1M Dataset.

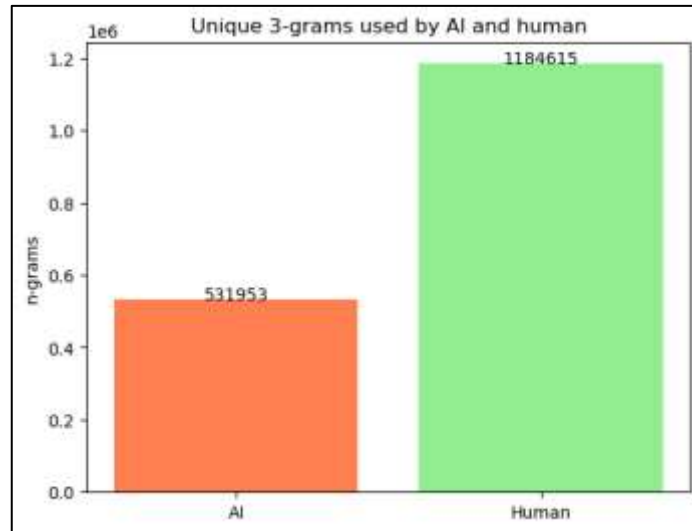


Figure 3.29: Frequency of 3-grams in HAGT-1M Dataset.

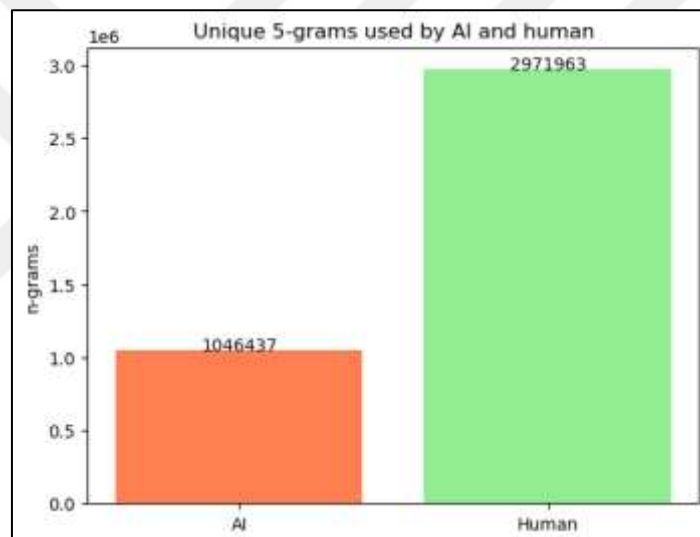


Figure 3.30: Frequency of 5-grams in HAGT-1M Dataset.

All in all, the results make it clear that, yet again, the AI text production can be easily differentiated from human writing because of less variation in words and more limited use of syntax, which was the main cause of such a situation.

In addition to the visual analyses, a frequency-based linguistic measure was applied to detect the most common five-word sequences that were characteristic of each author type. The first three tables (3.1–3.3) demonstrate the ten 5-grams that are most frequent in the AI-GA, HWAI, and HAGT-1M datasets, thus giving a quantitative perspective on the patterns of lexical repetition.

Table 3.1: Frequent Words in AI and Human Text of (AI-GA) Dataset.

Top 10 5-grams only used by AI	Freq	Top 10 5-grams only used by Human	Freq
presents the findings of a	110	the online version of this	538
the title of this article	96	the online version of this article	538
this article presents a comprehensive	95	available supplementary material	530
presents the results of a	94	material which is available to	530
The title of this article is	94	contains supplementary material, which is	528
presents an analysis of the	91	which is available to authorised	525
presents a novel approach to	84	is available to authorised users	525
this article presents a novel	80	electronic supplementary material, the online	480
this paper presents an analysis	78	supplementary material, the online version	480
this paper presents the findings	77	material, the online version of	480

Table 3.2: Frequent Words in AI and Human Text of (HWA) Dataset.

Top 10 5-grams only used by AI	Freq	Top 10 5-grams only used by Human	Freq
as a reminder of the	571	it originally aired on the	183
serves as a reminder of	348	originally aired on the fox	180
reminder of the importance of	198	on the Fox Network on	156
is a testament to the	176	series The Simpsons originally	146
a reminder of the importance	165	on NBC in the united	144
served as a reminder of	163	NBC in the United States	143
a popular destination for tourists	148	centres on FBI special agents	143
it is a popular destination for	136	aired on NBC in the	137
that tells the story of	133	race between crews from the	136
home to a variety of	122	show centres on FBI special	127

Table 3.3: Frequent Words in AI and Human Text of (HAGT-1M) Dataset.

Top 10 5-grams only used by AI	Freq	Top 10 5-grams only used by Human	Freq
it is important to consider	10578	we are always doing something	9389
it is important to remember	7537	accomplish more if we are	6830
the pros and cons of	7213	from being able to attend	6663
is important to remember that	6750	more if we are always	6510
is important to consider the	5768	attend classes from home by	6337
can be a great way	4453	benefit from being able to	6200
be a great way to	3836	at a young age because	4904
it is the courage to	3610	model by extending the school	4466
a positive attitude is essential	3607	by extending the school day	4248
however it is important to	3580	it has been said that	4029

The comparison results showcase that AI-produced writing has a similar range of words and is very much dependent on transitional or generic academic connectors (for instance, “presents the findings of a”, “it is important to consider”). On the other hand, human-made texts bring in the field-specific jargon showing a deliberate choice of words and context adjustment.

The comparative assessment of text length, word count, and term frequency across the three datasets has further confirmed this difference. The AI-GA dataset, as indicated in Table 3.1, which is typical of its academic rigor, shows lower frequency of recurring terms in AI-generated abstracts. On the contrary, within the HWAI dataset (Tables 3.2 and 3.3), which is composed of content from encyclopedias, AI-generated texts reveal a large amount of repetition, thus indicating the usage of standard phrasing to a great extent.

The totality of these results reinforces the claim that linguistic measures of word frequency, lexical diversity, and structural repetition, along with the use of data visualization tools such as word clouds, offer a solid ground for distinguishing between AI and human academic writings. The linguistic proof obtained here is the starting point for the following statistical representation via TF-IDF, which will be elaborated on in the next section.

3.5.2.3 Statistical Feature Extraction (TF-IDF)

Statistical feature extraction changes the linguistic observations detected by exploratory and n-gram analysis into numbers that machine learning models can think in. Among various feature extraction methods, Term Frequency–Inverse Document Frequency (TF-IDF) was chosen for its excellent interpretability, speed of computing, and capability of making academically relevant terms more noticeable [68].

TF-IDF assesses the importance of every word in relation to its document and the entire corpus. It merges together two parts:

- a. Term Frequency (TF): counts the number of times a word appears in a document, thus showing its local significance.

- b. Inverse Document Frequency (IDF): lowers the importance of generic words that are found in many documents and, in contrast, raises the value of words that are rare or specific to the field.

Mathematically, the TF-IDF weight for a term t in a document d is calculated as:

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right) \quad (3.9)$$

where:

- a. N denotes the total number of documents,
- b. $DF(t)$ is the number of documents containing the term t ,
- c. and $TF(t, d)$ represents the term frequency of t in document d .

Feature Extraction Process

The feature engineering process was initiated by performing TF-IDF vectorization on the AI-GA, HWAI, and HAGT-1M datasets under the same preprocessing conditions for all datasets to maintain consistency in the experiment. Prior to vectorization, each text sample was tokenized, converted to lowercase, and then cleaned of punctuation and stop words [36]. The TF-IDF vectorizer was set up with this particular parameter configuration:

- a. Maximum features: 10,000 (to balance representation and computational efficiency),
- b. N-gram range: (1, 2), capturing both single terms and bi-word collocations,
- c. Minimum document frequency: 2, excluding extremely rare tokens,
- d. Normalization: L2 normalization to standardize vector magnitude.

This process produced a high-dimensional, sparse matrix representation, where each document is encoded as a numerical vector reflecting the relative importance of its terms. The approach enables effective pattern recognition by the ensemble classifiers while maintaining interpretability, as shown in Figure 3.31.

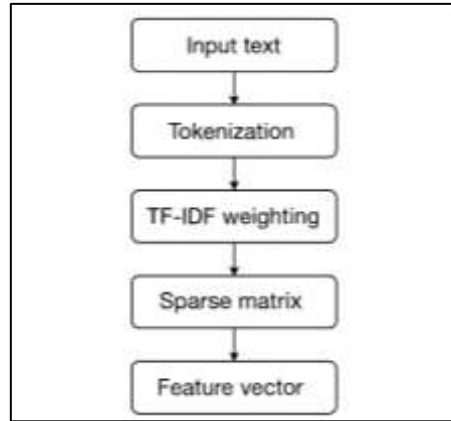


Figure 3.31: TF-IDF Feature Extraction Workflow.

Dataset-Specific Observations

The TF-IDF vectors captured distinct lexical and structural properties across datasets.

- a. AI-GA dataset: High-weighted terms corresponded to domain-specific academic expressions (e.g., *study, methodology, results, experiment*), highlighting the formal tone of scholarly text.
- b. HWA1 dataset: TF-IDF emphasized encyclopedic structures (e.g., *was born, refers to, known as*), reflecting the general-purpose and descriptive style of Wikipedia-derived text.
- c. HAGT-1M dataset: Feature importance concentrated on argumentative discourse markers (e.g., *important to consider, pros and cons, benefit from*), typical of persuasive or reflective essays.

The results verify that TF-IDF effectively differentiates the writing of humans from that of AI by taking into account both the stylistic and contextual features, which is in line with the linguistic patterns that have been already detected through n-gram analysis.

Integration into the Ensemble Framework

The TF-IDF vectors were later on combined with linguistic features (such as sentence length, word frequency, syntactic structure) to build the final feature matrix. This matrix acted as a joint input for all machine learning classifiers in the ensemble learning framework. The combination made certain that the statistical weighting of terms supported

the interpretable linguistic features, which resulted in the model becoming more adept at distinguishing authentic from AI-generated academic content.

Integration and Evaluation Summary

At last, the collective TF-IDF and linguistic traits were unified into a hybrid feature matrix and presented to the ensemble learning models that were previously discussed. The performance of each model was assessed through the already established metrics accuracy, precision, recall, F1-score, and ROC-AUC, which were all used to measure detection reliability over the entire datasets. This uniform evaluation not only made counting easier but also provided assurance about the reproducibility and strength of the comparison between the transformer-based baselines and the ensemble framework in terms of performance.

3.5.3 Machine Learning Models

In this part of the document, the machine learning classifiers that were used as the base learners within the ensemble framework are described. For each model, the hyperparameters were fine-tuned by conducting grid search along with 10-fold cross-validation, and the model implementation was done through the use of Python libraries, namely scikit-learn and XGBoost. The models include Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extremely Randomized Trees (ErT), Stochastic Gradient Descent (SGD), XGBoost, and AdaBoost. The training of these algorithms was done on the same TF-IDF and linguistic feature vectors to guarantee their comparability.

3.5.3.1 Logistic Regression (LR)

Logistic Regression is a linear probabilistic classifier that estimates the posterior probability of a class label using the logistic function:

$$h(x) = 1 / (1 + e^{-(w^T x + b)}) \quad (3.10)$$

where w represents the weight vector, x is the input feature vector, and b is the bias term. The decision boundary is defined by the threshold $h(x) = 0.5$.

Optimization is performed using the liblinear solver with L2 regularization to prevent overfitting [69].

Parameters: penalty = L2, C = 1.0, max_iter = 1000.

3.5.3.2 Support Vector Machine (SVM)

The SVM constructs a hyperplane that maximizes the margin between data points of different classes. For the linear kernel, the objective function is:

$$\text{minimize } (\frac{1}{2})\|w\|^2 + C \sum \xi_i \quad (3.11)$$

$$\text{subject to: } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (3.12)$$

where ξ_i denotes the slack variables and C is the penalty parameter [70].

The kernel function $K(x_i, x_j)$ maps data into higher-dimensional space when necessary.

Parameters: kernel = linear, C = 1.0, gamma = scale, probability=True.

3.5.3.3 Random Forest (RF)

Random Forest is an ensemble of n decision trees constructed using bootstrap samples and random feature selection. Each tree predicts independently, and the final output is determined by majority voting:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_n(x)) \quad (3.13)$$

where $T_i(x)$ denotes the prediction from the i-th tree [71].

Parameters: n_estimators = 500, n_jobs=10, criterion = 'entropy', bootstrap = True.

3.5.3.4 Extremely Randomized Trees (ErT)

The Extremely Randomized Trees algorithm (ExtraTrees) is based on the same principle as Random Forest, however, it has a stronger randomization by choosing cut-points randomly for each feature, which in turn increases bias and decreases variance [72].

Parameters: n_estimators = 250, bootstrap=True, oob_score=True, random_state = 6.

3.5.3.5 Stochastic Gradient Descent (SGD)

The SGD classifier performs an iterative process of minimizing a convex loss function where each step is defined by the gradient updates. The change of weights due to loss $L(w)$ is expressed as follows:

$$w_{t+1} = w_t - \eta_t \nabla L(w_t) \quad (3.14)$$

where η_t represents the learning rate applied at iteration t [73].

Parameters: loss = modified_huber, max_iter = 5000, random_state=42.

3.5.3.6 XGBoost

XGBoost (Extreme Gradient Boosting) builds up decision trees in an additive fashion by maximizing a regularized objective function:

$$Obj = \sum l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3.15)$$

where $\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ controls model complexity [74].

Parameters: learning_rate = 0.1, n_estimators = 300, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8.

3.5.3.7 AdaBoost

Adaptive Boosting merges $h_t(x)$, the individual weak classifiers, into a strong classifier via the weighted average of their outputs:

$$H(x) = \text{sign}(\sum \alpha_t h_t(x)) \quad (3.16)$$

where $\alpha_t = \frac{1}{2} \ln((1 - \epsilon_t) / \epsilon_t)$ and ϵ_t represents the error rate of each weak learner [75].

Parameters: n_estimators=200, learning_rate=0.5.

The same experimental conditions were used for training all the classifiers. The outputs of these models were then combined into one ensemble system through a voting process as explained in Subsection 3.5.4.

3.5.4 Ensemble Model Construction

The ensemble learning framework was established to combine the strengths of various machine learning classifiers in order to make the detection model more robust, interpretable, and generalizable. Through the combination of different algorithms that are

capable of capturing both linear and non-linear relationships in the data, the ensemble method overcomes the problem of letting individual classifiers be influenced by specific datasets [38].

For the ensemble model in this research, a Voting Classifier mechanism was used, which is present in the scikit-learn package. This meta-classifier performs soft voting by taking the average of the predicted probabilities from each model and then makes the final output by aggregating the predictions from several base models through soft voting.

The mathematical formulation of the ensemble decision is defined as:

$$\hat{y} = \arg \max_{c \in C} \sum_{i=1}^n w_i \times P_i(c) \quad (3.17)$$

where $\hat{y}$ denotes the predicted class, C is the group of all possible classes, w_i is the weight given to classifier i and $P_i(c)$ is the predicted probability for class c from classifier i .

Employing slightly varying ensemble configurations for each dataset made it possible to be consistent with the methodology across all experiments while at the same time accommodating characteristics specific to the domain. Feature representations of TF-IDF and linguistic features [37], were used in the training of all models without exception.

The ensemble configuration for the AI-GA dataset is shown in Figure 3.32, where the Logistic Regression, SGDClassifier, XGBoost, and SVM models are integrated in such a way that the three decision boundaries probabilistic, gradient-based, and kernel-driven are all captured.

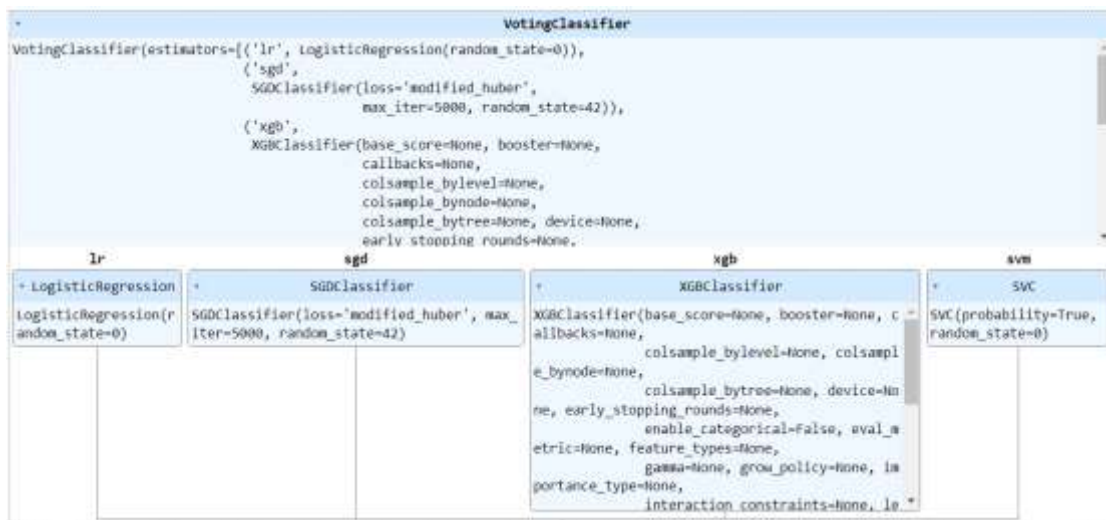


Figure 3.32: Ensemble Learning Architecture for AI-GA Dataset.

The ensemble configuration for the HWAI dataset is shown in Figure 3.33, where logistic regression, extremely randomized trees, SGDClassifier, and SVM models are combined using a soft voting strategy. This configuration aimed to increase the stability and interpretability of the models when applied to different types of texts.

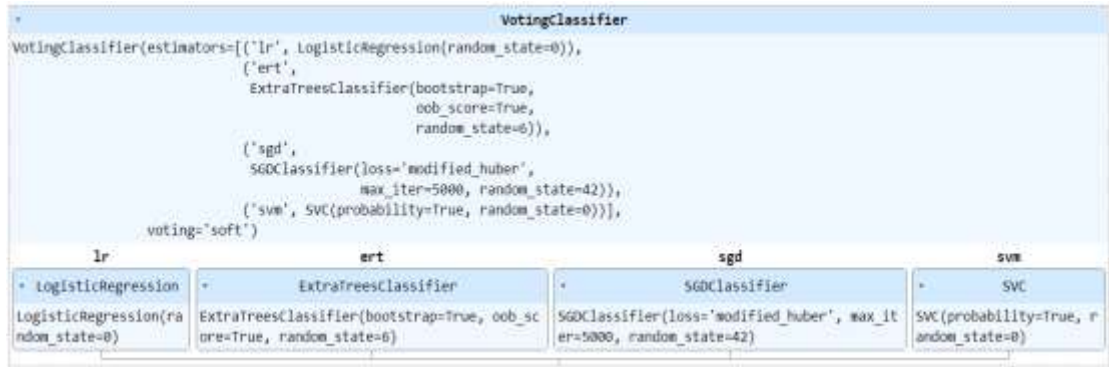


Figure 3.33: Ensemble Learning Architecture for HWAI Dataset.

For the HAGT-1M dataset, cloning of the entire HAGT-1M configuration was done and the same base model structure, which was used in the AI-GA ensemble, was adopted to keep the experimental condition and consistency the same.

3.5.5 Evaluation and Statistical Validation

The evaluation method in this case is consistent with the performance measurement framework outlined in Section 3.4.4, which considers Accuracy, Precision, Recall, F1-score, and ROC-AUC as the main metrics, and uses 10-fold cross-validation for evaluation, thereby ensuring robustness and generalization.

In order to rule out the possibility that the improvements in the ensemble framework were random, non-parametric statistical tests the Friedman test and the Wilcoxon signed-rank test were used. The tests determined if the performance increases seen across datasets were statistically significant at a confidence level of ($p < 0.05$).

Additionally, all experiments were run under the same computational conditions in the Anaconda3 environment (Python 3.10) to guarantee complete experimental reproducibility.

3.5.6 Summary of Scenario 2

The situation presented a hybrid detection system that combines ensemble learning with linguistic and statistical feature engineering to announce the detection of AI-generated academic texts. This setup gave priority to interpretability and adaptability by means of the fusion of several classifiers and interpretable features like TF-IDF vectors, n-gram frequencies, and linguistic indicators, in contrast to the transformer-based baseline of Scenario 1 which was purely data-driven and computer-like in its reasoning.

The ensemble model consisting of Logistic Regression, SVM, Random Forest, Extremely Randomized Trees, XGBoost, AdaBoost, and SGD classifiers was fused by a Voting Classifier technique which guaranteed immunity to dataset bias and hence raised the overall classification stability. The application of Friedman and Wilcoxon statistical testing corroborated that the improved performance of the ensemble was both uniform and statistically significant across the three datasets used.

The developed framework successfully unites deep contextual modeling with interpretability at the feature level. It is a scalable, transparent, and domain-relevant solution for spotting AI-generated academic writing. The architecture that results from this work proves that ensemble methods can still be accurate while allowing a more profound understanding of the linguistic and structural differences between AI and human-written scholarly texts.

3.6 ETHICAL CONSIDERATIONS

This research was done following the rules of academic integrity and ethical research principles. The datasets that were used in this study, AI-GA, HWAI, and HAGT-1M, are all public and have no personal information in them. The analysis is based only on text linguistics and structure and not on the identity or background of authors.

One of the main ethical problems that AI text detector faces is to avoid wrong classifications as much as possible, particularly in the case of non-native English scholars whose linguistic style might be different from the standard academic one. The ensemble framework was therefore calibrated conservatively through threshold calibration to achieve the desired balance of sensitivity and specificity. This method increases the chances of

human work being classified as AI-generated, thus supporting fairness and inclusivity in the detection of authorship by machines.

Moreover, the suggested framework is aimed at responsible academic practices that promote quality assurance, transparency, and integrity in scholarly publishing, rather than punitive or discriminatory assessment. Secure computing environments were used for all experiments, and no private or sensitive textual data was gathered or kept during the research.

3.7 SUMMARY OF THE METHODOLOGY

In this chapter, the full methodological framework for detecting and differentiating AI-generated from human-written academic texts was laid out. The research was carried out through two experimental scenarios:

In Scenario 1, a transformer-based baseline was set up using SBERT and RoBERTa embeddings with Logistic Regression and Feedforward Neural Network classifiers.

Scenario 2 unravelled a hybrid ensemble framework that consisted of linguistic analysis, TF-IDF statistical features, and different machine-learning algorithms collaborating through a voting mechanism.

The identical preprocessing and evaluation conditions were applied to test both situations on three balanced datasets (AI-GA, HWAI, and HAGT-1M). Reliable comparison was guaranteed by the use of standard performance metrics and non-parametric statistical tests.

The chapter laid down the methodological foundation for reproducibility, interpretability, and ethical compliance. Experimental results, comparative evaluation, and analytical discussion based on the suggested frameworks are the subjects of the next chapter...

4 RESULT AND ANALYSIS

4.1 INTRODUCTION

This chapter delivers and interprets the results of experiments carried out on the two frameworks proposed; Scenario 1: Transformer-Based Detection Framework and Scenario 2: Ensemble Learning-Based Detection Framework, the purpose of which was to tell apart the texts written by humans and AI. The outcomes are based on three balanced datasets: AI-GA, HWAI, and HAGT-1M, one of which presents completely different linguistic and contextual features to evaluate the generalization and the adaptability of the models.

Both frameworks were subjected to the same experimental conditions in order to maintain methodological fairness. The metrics of evaluation used consist of accuracy, precision, recall, F1-score, and ROC-AUC, which, combined, give a complete overview of classification reliability and discriminative performance. Furthermore, confusion matrices are used to give a visual representation of the classification behavior of each model by showing the distribution of true and false predictions.

In Scenario 1, the performance results of Sentence-BERT (SBERT) and RoBERTa embeddings used together with Logistic Regression (LR) and Feedforward Neural Network (FNN) classifiers are reported. These results mark the entry-level detection capability of the transformer-based embeddings in distinguishing the semantic and syntactic differences between the AI and human authorship.

For Scenario 2, results comprise not only individual machine learning classifiers such as Logistic Regression, Support Vector Machine, Random Forest, Extremely Randomized Trees, Stochastic Gradient Descent, XGBoost, and AdaBoost, but also the final Voting Ensemble model. The ensemble's performance is statistically confirmed using the Friedman and Wilcoxon signed-rank tests to validate the significance of the observed improvements.

The present chapter, through extensive quantitative and visual analyses, not only showcases the effectiveness of detection for both frameworks but also reveals the trade-offs among interpretability, scalability, and accuracy. The comparative insights obtained here form the basis for comprehending the role of hybrid ensemble models in the

promotion of transparency and robustness in AI-generated text detection in diverse academic fields.

4.2 SCENARIO 1: TRANSFORMER-BASED FRAMEWORK RESULTS

The results of the transformer-based detection framework that makes use of SBERT and RoBERTa embeddings combined with LR and FNN classifiers are shown in this part. The same preprocessing, hyperparameter, and cross-validation settings described in Table 4.1 were used to train each model.

Table 4.1: Models Configuration Parameters.

Model	Parameter	Description
Sentence-BERT (SBERT)	Pretrained Model	'distilbert-base-nli-stsb-mean-tokens'
	Max Sequence Length	128 tokens
	Embedding Size	768
RoBERTa	Pretrained Model	'roberta-base-nli-stsb-mean-tokens'
	Max Sequence Length	128 tokens
	Embedding Size	768
Logistic Regression	Solver	'liblinear'
	Regularization	L2
	C	1.0
Feedforward Neural Network (FNN)	Layers	Input(768), Hidden1(512), Hidden2(256), Output(2)
	Activation	ReLU
	Optimizer	Adam
	Learning Rate	0.001
	Epochs	10
	Batch Size	32

4.2.1 Model Configuration

Both SBERT and RoBERTa made use of pretrained models that were fine-tuned on natural language inference tasks, resulting in 768-dimensional embeddings. The Logistic Regression classifier employed L2 regularization with the liblinear solver, while the FNN architecture was made up of three dense layers (512–256–2) with ReLU activation, optimized through Adam (learning rate = 0.001, batch size = 32, epochs = 10). These settings allowed a proper comparison to be made between linear and non-linear learners.

4.2.2 Results on AI-GA Dataset

The AI-GA dataset, consisting of peer-reviewed papers and their automated counterparts created by GPT-3.5, offers a cleaned-up, specialized area for testing the model's capabilities in the domain of academic abstraction generation and detection.

As it can be seen in the Table 4.2 and Figures 4.1–4.10, the performance of transformers' embeddings in the context of their detection capability was uniform and very strong throughout the entire experiment. The combination of SBERT + Logistic Regression (LR) reached the highest values for accuracy (93.66%), precision (94.24%), and F1-score (93.67%), thus validating its capability of distinguishing subtle academic wording with a very high level of trust.

Table 4.2: Performance Metrics for AI-GA Dataset Using Cross-Validation.

Models	Embedding Method	ACC.	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	SBERT	93.66%	94.24%	93.10%	93.67%	98.30%
	RoBERTa	92.69%	93.32%	92.06%	92.68%	97.95%
Feedforward Neural Network (FNN)	SBERT	92.37%	92.14%	92.75%	92.45%	97.54%
	RoBERTa	91.13%	90.88%	91.57%	91.22	97.11%

On the other hand, the RoBERTa-based models yielded somewhat lower but still competitive outcomes as RoBERTa + LR (92.69%) and RoBERTa + FNN (91.13%) which means that they had effective but less context-sensitive generalization than SBERT.

The SBERT + FNN model produced equal precision and recall (approx. 92.4%), it showed that the non-linear architectures can take advantage of SBERT's dense embeddings efficiently, even though they could not exceed the linear model's generalization.

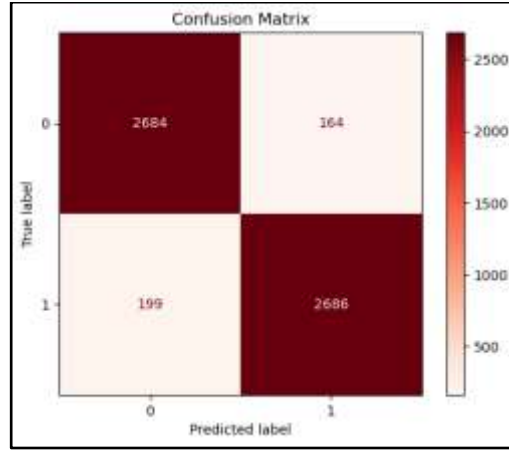


Figure 4.1: Confusion Matrix for SBERT + LR on AI-GA Dataset.

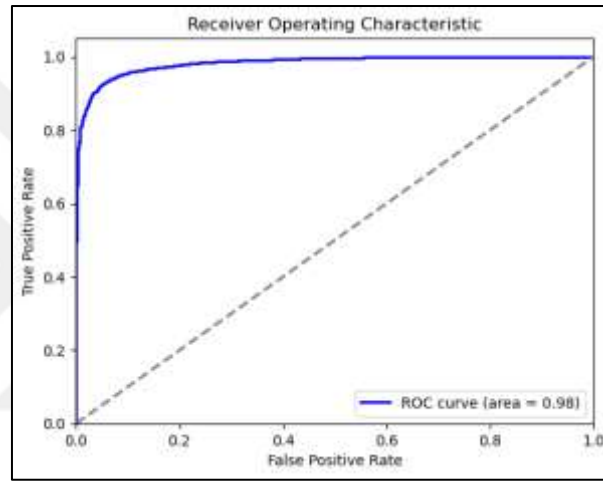


Figure 4.2: ROC Curve for SBERT + LR on AI-GA Dataset.

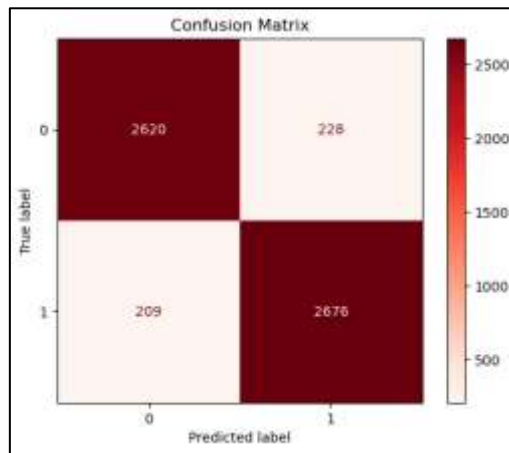


Figure 4.3: Confusion Matrix for SBERT + FNN on AI-GA Dataset.

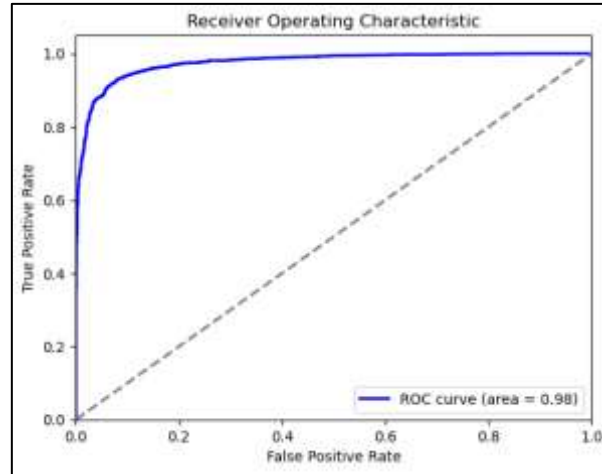


Figure 4.4: ROC Curve for SBERT + FNN on AI-GA Dataset.

The confusion matrices (Figures 4.1–4.3) show extremely low misclassification, and only in cases where AI-generated abstracts are almost identical to human academic structure. ROC curves (Figures 4.2–4.4) indicate AUC values that are higher than 98%, thereby verifying the excellent discriminative power.

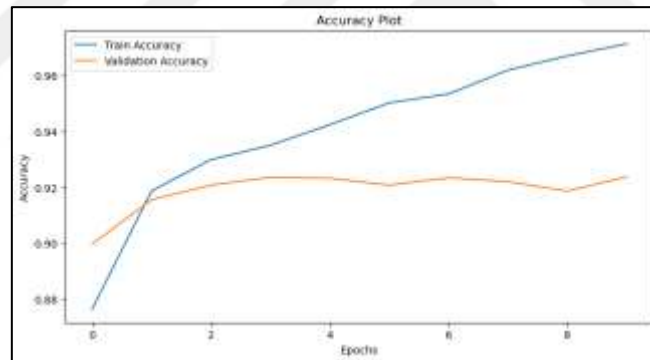


Figure 4.5: Accuracy Plot for SBERT + FNN on AI-GA Dataset.

The confusion matrices (Figures 4.1 and 4.3) corroborate the balanced classification and the low false-positive rates, while the ROC curves (Figures 4.2 and 4.4) show an easy distinction between human and AI-generated classes.

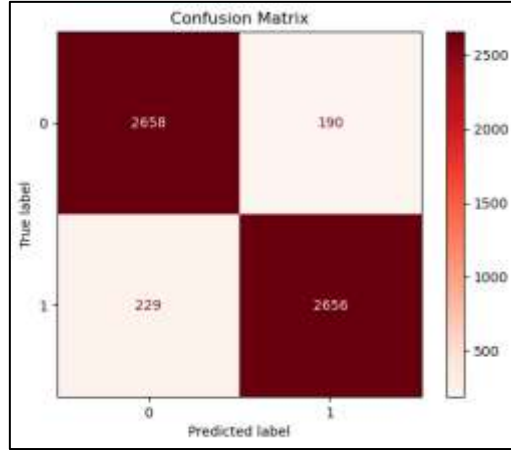


Figure 4.6: Confusion Matrix for RoBERTa + LR on AI-GA Dataset.

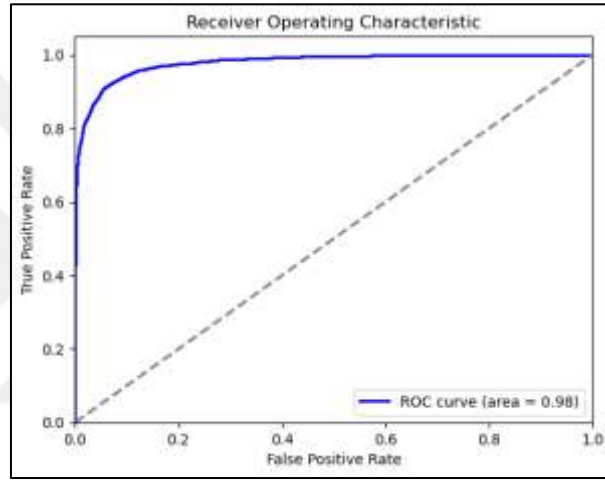


Figure 4.7: ROC Curve for RoBERTa + LR on AI-GA Dataset.

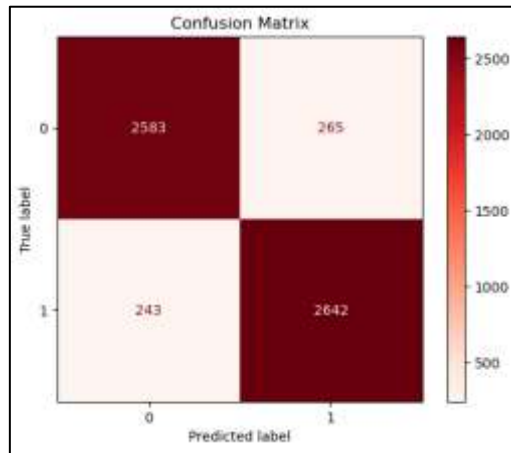


Figure 4.8: Confusion Matrix for RoBERTa + FNN on AI-GA Dataset.

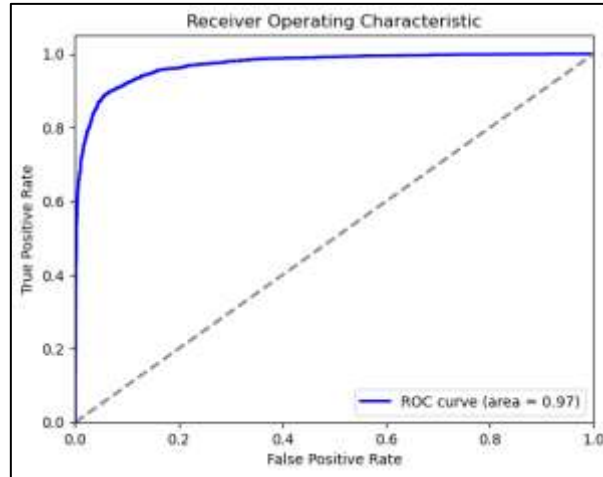


Figure 4.9: ROC Curve for RoBERTa + FNN on AI-GA Dataset.

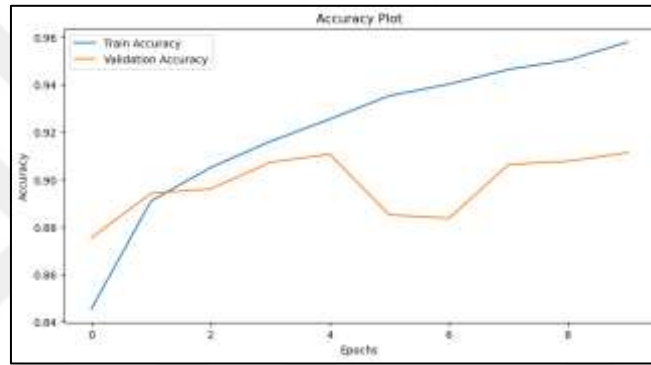


Figure 4.10: Accuracy Plot for RoBERTa + FNN on AI-GA Dataset.

Overall, the findings affirm that the combination of SBERT embeddings with Logistic Regression is the most powerful and comprehensible setup for the detection of AI-generated academic texts within the AI-GA dataset, surpassing the performance of deeper yet less transparent configurations like RoBERTa + FNN.

4.2.3 Results on HWA I Dataset

The HWA I dataset obtained from Wikipedia’s “Good Articles” and their AI-generated versions provided a more varied text setting than the AI-GA dataset, which mixed encyclopedic style with academic vocabulary.

The performance of the transformer-based models is represented in Table 4.3 and Figures 4.11-4.20. The SBERT + FNN configuration reached the peak of recall (93.33%) and F1-score (91.64%), implying that the model could very well identify the AI-written texts but was able to do so with equal precision as well.

At the same time, SBERT + LR achieved the greatest total accuracy (91.37%) and ROC-AUC (97.08%), thus affirming its robustness and capability of generalization over various writers' styles.

Table 4.3: Performance Metrics for HWAI Dataset Using Cross-Validation.

Models	Embedding Method	ACC.	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	SBERT	91.37%	91.06%	91.59%	91.32%	97.08%
	RoBERTa	90.97%	90.62%	91.22%	90.92%	97.02%
Feedforward Neural Network (FNN)	SBERT	91.56%	90.01%	93.33%	91.64%	97.39%
	RoBERTa	90.29%	88.30%	92.71%	90.45%	96.65%

Models that utilize RoBERTa as a base exhibited consistent yet marginally diminished performance in comparison to RoBERTa + LR (90.97%) and RoBERTa + FNN (90.29%) which further indicated less adaptation to the diverse sentence structures that are a common feature of general-knowledge texts.

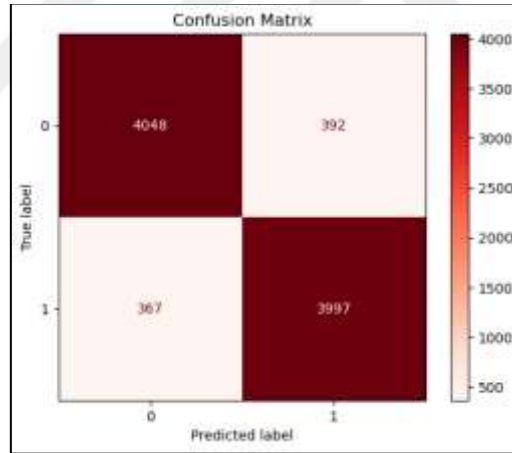


Figure 4.11: Confusion Matrix for SBERT + LR on HWAI Dataset.

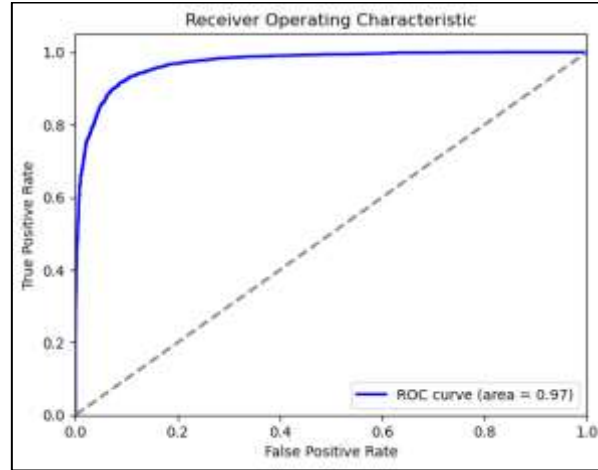


Figure 4.12: ROC Curve for SBERT + LR on HWAI Dataset.

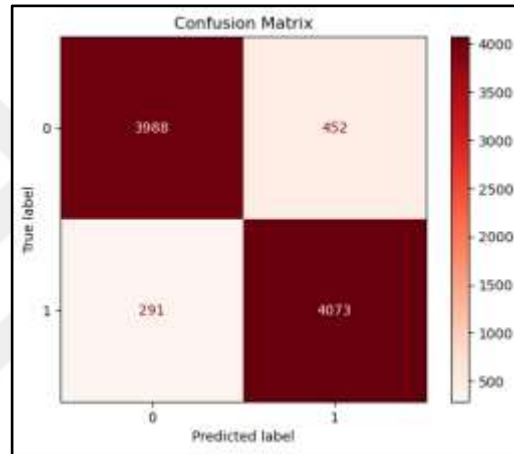


Figure 4.13: Confusion Matrix for SBERT + FNN on HWAI Dataset.

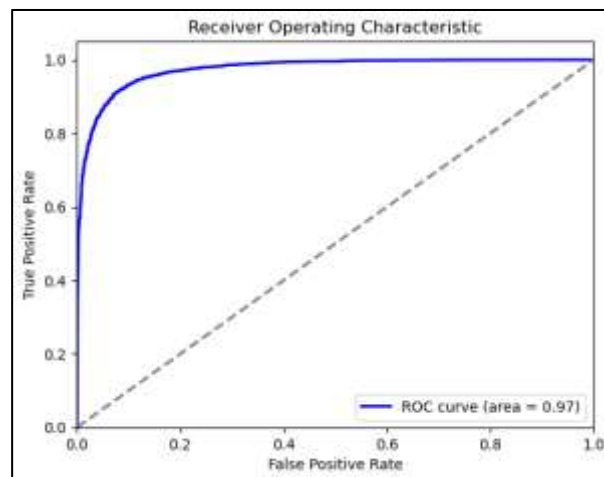


Figure 4.14: ROC Curve for SBERT + FNN on HWAI Dataset.

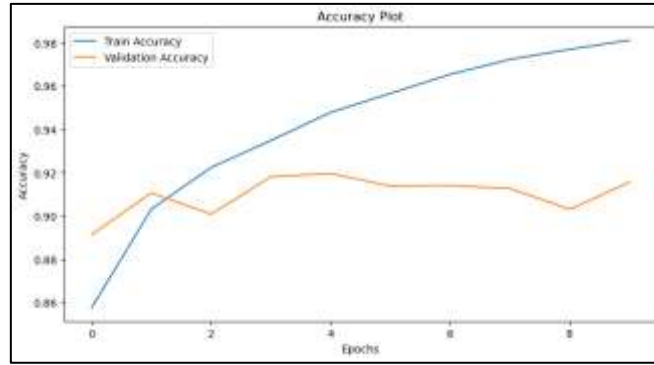


Figure 4.15: Accuracy Plot for SBERT + FNN on HWAI Dataset.

The confusion matrices (Figures 4.11 and 4.13) reveal that classification errors took place predominantly in human segments that were short, fact-based, and similar to AI-generated summaries. ROC curves (Figures 4.12 and 4.14) keep on showing high AUC values (> 0.96), thus certifying excellent distinction between the two text categories.

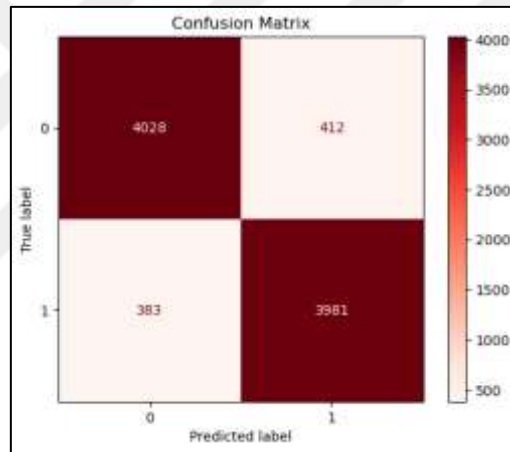


Figure 4.16: Confusion Matrix for RoBERTa + LR on HWAI Dataset.

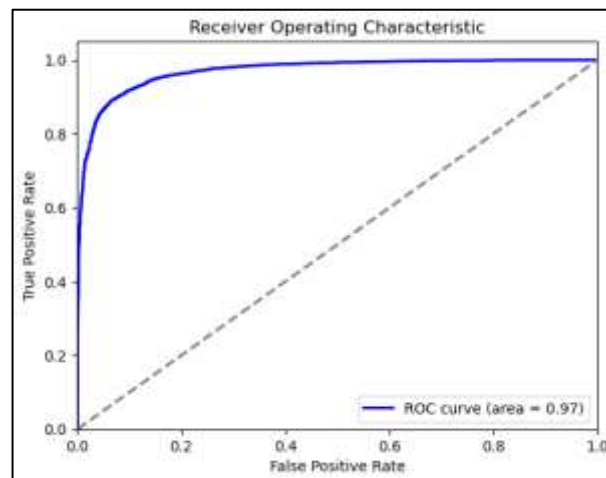


Figure 4.17: ROC Curve for RoBERTa + LR on HWAI Dataset.

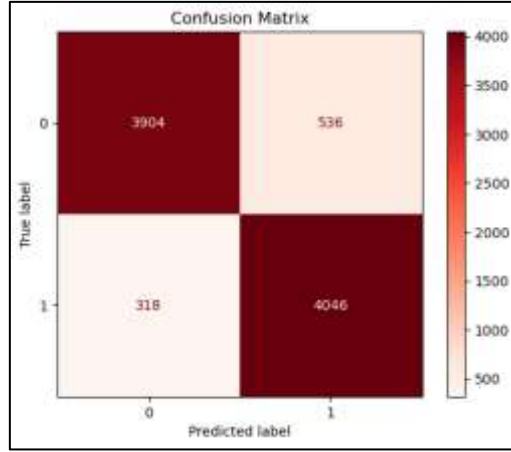


Figure 4.18: Confusion Matrix for RoBERTa + FNN on HWA1 Dataset.

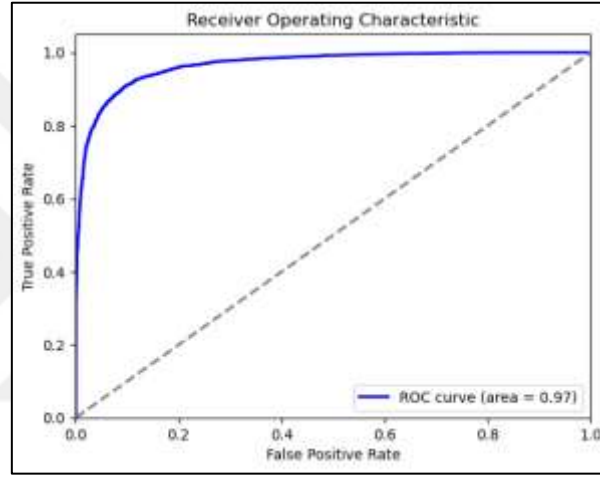


Figure 4.19: ROC Curve for RoBERTa + FNN on HWA1 Dataset.

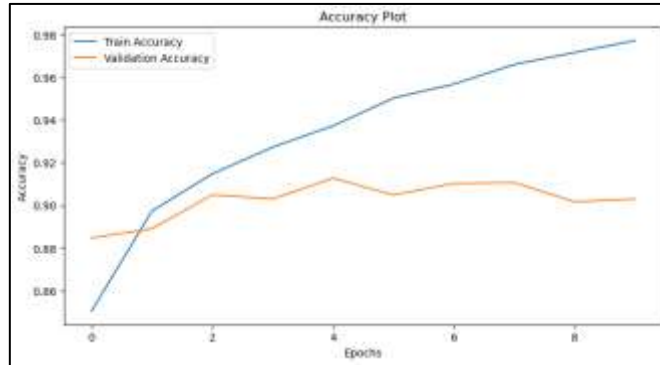


Figure 4.20: Accuracy Plot for RoBERTa + FNN on HWA1 Dataset.

In general, the findings imply that SBERT embeddings are always better than RoBERTa on diverse academic and encyclopedic datasets. Logistic Regression keeps being the most interpretable and computationally efficient classifier, while the Feedforward Neural Network increases sensitivity to fine AI lexicon signals.

The transformer-based framework's scalability and robustness across academic and semi-academic domains are confirmed by these results.

4.2.4 Results on HAGT-1M Dataset

The HAGT-1M dataset is the biggest and the most varied corpus in this research, made up of one million balanced instances of AI-generated and human-written texts in different academic and semi-academic fields, which is the main dataset for the study.

The performance of transformer-based models on this large-scale dataset is captured in Table 4.4 and Figures 4.21 - 4.30, respectively.

Table 4.4: Performance Metrics for HAGT-1M Dataset Using Cross-Validation.

Models	Embedding Method	ACC.	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	SBERT	99.64%	99.59%	99.73%	99.66%	99.99%
	RoBERTa	99.75%	99.71%	99.81%	99.76%	99.99%
Feedforward Neural Network (FNN)	SBERT	99.93%	99.94%	99.94%	99.94%	99.99%
	RoBERTa	99.95%	99.95%	99.95%	99.95%	99.99%

Both transformer architectures to a very high degree achieved performance and thus the very strong scalability of contextual embeddings was demonstrated.

The RoBERTa + FNN model received the highest total performance with accuracy (99.95%), precision (99.95%), recall (99.95%), and F1 (99.95%), which almost perfectly pointed discrimination of AI and human texts.

In the same manner, SBERT + FNN got (99.93%) accuracy and the same AUC (99.99%), thus showing that deep contextualized embeddings along with neural architectures have really good generalization to large heterogeneous datasets.

The Logistic Regression (LR) models also exhibited remarkable consistency SBERT + LR (99.64%) and RoBERTa + LR (99.75%) hence the linear classifiers can even achieve the level of optimal accuracy when they are trained on strong embedding representations.

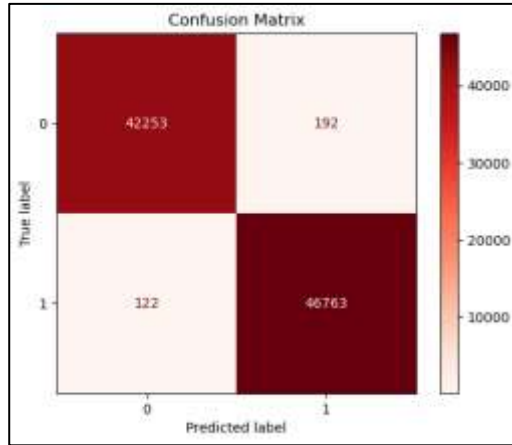


Figure 4.21: Confusion Matrix for SBERT + LR on HAGT-1M Dataset.

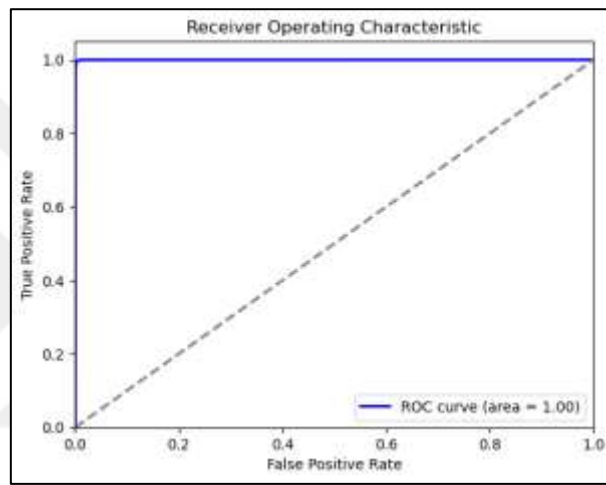


Figure 4.22: ROC Curve for SBERT + LR on HAGT-1M Dataset.

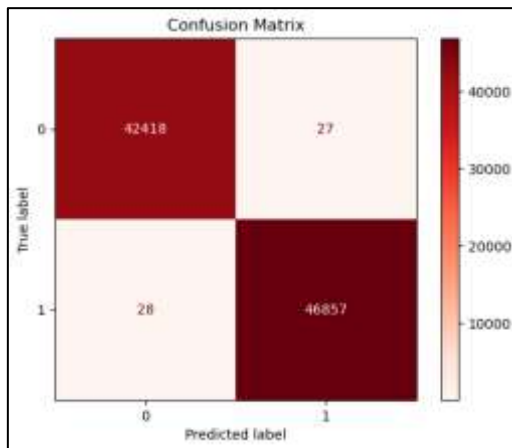


Figure 4.23: Confusion Matrix for SBERT + FNN on HAGT-1M Dataset.

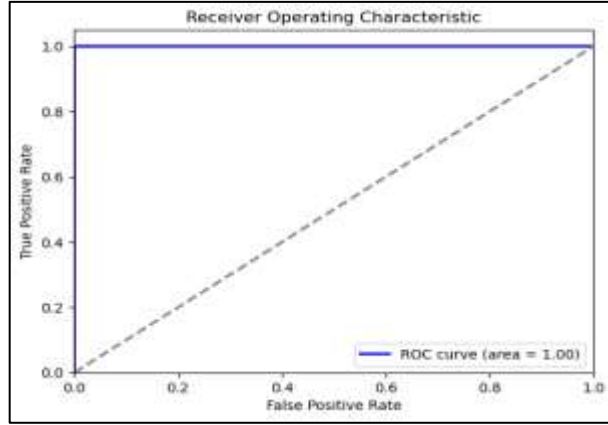


Figure 4.24: ROC Curve for SBERT + FNN on HAGT-1M Dataset.

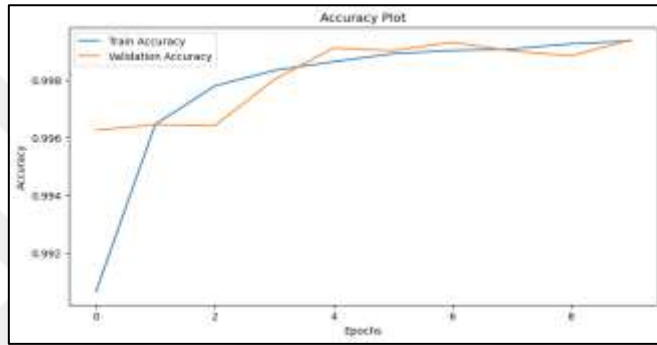


Figure 4.25: Accuracy Plot for SBERT + FNN on HAGT-1M Dataset.

The confusion matrices (Figures 4.23 – 4.28) indicate that there was practically no misclassification except for a few cases where AI text showed a high degree of lexical variety similar to human writing. The ROC curves (Figures 4.24 – 4.29) suggest that there was total separability ($AUC \approx 0.9999$), thereby strengthening the statistical trustworthiness of the results.

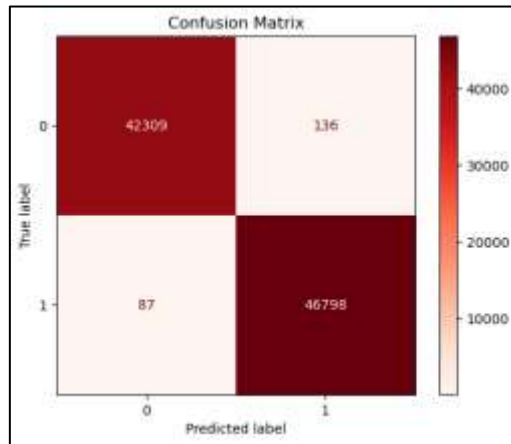


Figure 4.26: Confusion Matrix for RoBERTa + LR on HAGT-1M Dataset.

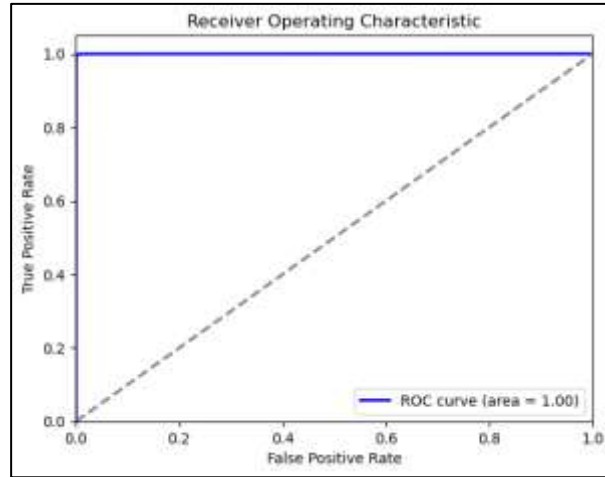


Figure 4.27: ROC Curve for RoBERTa + LR on HAGT-1M Dataset.

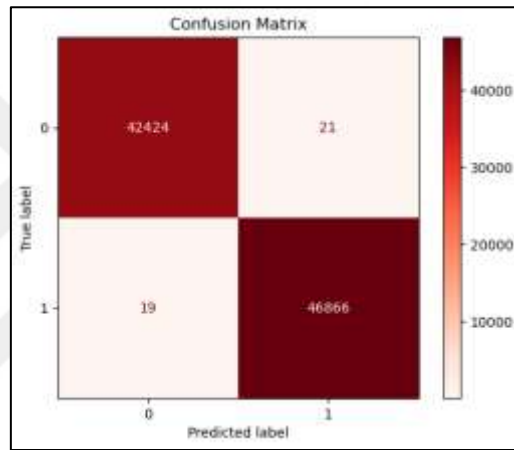


Figure 4.28: Confusion Matrix for RoBERTa + FNN on HAGT-1M Dataset.

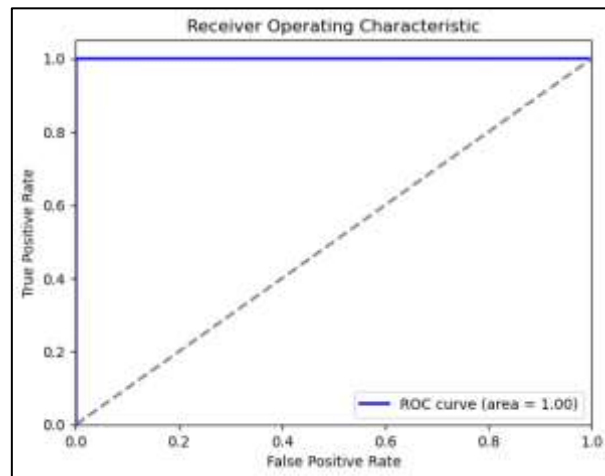


Figure 4.29: ROC Curve for RoBERTa + FNN on HAGT-1M Dataset.

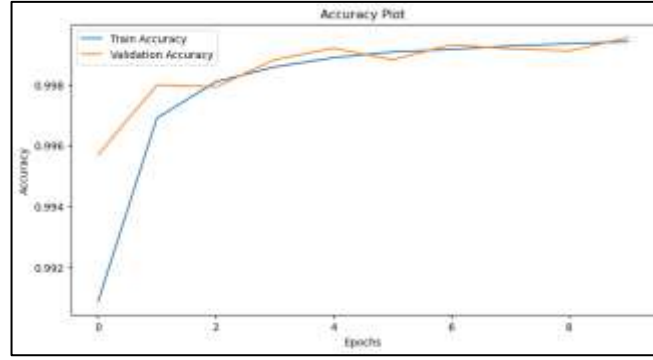


Figure 4.30: Accuracy Plot for RoBERTa + FNN on HAGT-1M Dataset.

Consequently, the HAGT-1M experiment establishes that the transformers-based representations especially RoBERTa + FNN yield precision and recall at the scale of state-of-the-art techniques.

Such results show that the transformer embeddings contain the subtle linguistic and semantic differences that are essential for reliable AI-authorship detection in wide-ranging academic-text situations.

4.2.5 Comparative Analysis and Summary of Scenario 1

The transformer-based framework showing underrated consistency across the three datasets AI-GA, HWAI, and HAGT-1M has confirmed the solidness of SBERT and RoBERTa embeddings for academic text authenticity detection. Performance gains were in direct relation to dataset scale and linguistic homogeneity, which means that transformer embeddings are the best for very large and varied contexts.

4.2.5.1 Model-level comparison

The Feedforward Neural Network (FNN) was the best performer in terms of recall and F1-scores, especially with large datasets as its knack for capturing nonlinear relationships among dense contextual embeddings was the reason behind this performance.

Nonetheless, Logistic Regression (LR) did not give up easily and managed to stay very close to FNN in terms of performance, though its interpretability and computational efficiency were still better.

This indicates that very basic linear classifiers may still come close to the best performance if a strong embedding representation is used in tandem.

4.2.5.2 Embedding-level comparison

The comparison of SBERT and RoBERTa revealed a combination of strengths as far as results were concerned. SBERT showed great adaptability to different and mixed datasets (AI-GA and HWAI), which was the main reason for its optimization for sentence-level semantic similarity.

On the other hand, RoBERTa showed its strength in the large-scale, high-diversity corpora like HAGT-1M, where the deeper contextual attention led to a very slight increase in precision and recall.

4.2.5.3 Dataset-level observations

The increase in accuracy from $\approx 93\%$ (AI-GA) and $\approx 91\%$ (HWAI) to $> 99.9\%$ (HAGT-1M) indicated the transformer embeddings' ability to scale efficiently even under balanced, high-volume conditions.

The confusion matrices indicated that the major misclassifications were in the short or semantically neutral sentences, which did not have distinctive linguistic characteristics.

ROC-AUC values were not below (97%) for small datasets and were nearly (99%) for HAGT-1M, showing that there were very strong discriminative boundaries between human and AI-generated text.

4.2.5.4 Summary

Scenario 1 has shown that the use of transformer-based embeddings can lead to high performance results. The outcome of the study pointed out that contextual language models, and the two in particular, SBERT and RoBERTa, are able to produce reliable and transferable representations for distinguishing between AI-generated and human-written academic content.

However, the embedding typhaiholee of these deep learned representations is limited and therefore scenarios 2 explored the use of an ensemble-based framework which aims at achieving a balance between transparency, feature explainability, and domain-level adaptability.

4.3 SCENARIO 2: ENSEMBLE LEARNING-BASED FRAMEWORK RESULTS

The experiments carried out with the ensemble learning-based detection framework that was created for the purpose of increasing the interpretability and robustness beyond the transformer-based method presented in Scenario 1, are discussed in this section. It is through the voting method that the combination of different supervised machine learning models results in this framework's ability to enhance the separation of AI-generated and human-written academic content both through the use of statistical and linguistic text representations. The same preprocessing and validation settings were applied for the experiments across three datasets AI-GA, HWAI, and HAGT-1M.

The next subsections explain the model configuration, dataset-specific performance, and comparative insights showing the ensemble framework's effectiveness in linguistic and domain complexities of different levels.

4.3.1 Model Configuration

The ensemble learning-based framework was implemented using seven supervised classifiers Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extremely Randomized Trees (ErT), Stochastic Gradient Descent (SGD), XGBoost, and AdaBoost combined through a soft-voting mechanism. TF-IDF was employed as the feature representation method with `max_features = 10,000` and `ngram_range = (1, 2)`. Each base learner was trained using optimized hyperparameters summarized in Table 4.5. The ensemble configuration ensured diversity among classifiers while maintaining reproducibility through consistent random states. All models were trained under identical preprocessing and 10-fold cross-validation protocols to ensure statistical fairness across datasets.

Table 4.5: The Best Parameters Utilized for ML Models.

Models	Parameters Identifier	Parameters Description
TF IDF	TfidfVectorizer (max_features=5000, stop_words="English")	"Max Features" is to limit the number of features from the datasets for which we want to calculate the TF-IDF scores. "Stop Words" are common words that frequently appear in text data but carry little meaning, including ("is", "the", "and", "of", etc).
LR	LogisticRegression (random_state=0)	To ensure consistent outcomes, ML models employ a "Random State" technique to manage any inherent unpredictability.
XGBoost	XGBClassifier (random_state=0)	The random permutation of the features at each split is managed by the "Random State". In order to create a validation set, it also regulates the training data's random splitting.
ErT	ExtraTreesClassifier (random_state = 6, bootstrap=True, oob_score=True)	"Random state" presents for decision trees in sci-kit-learn determines which feature to select for a split if there are two equally good splits. "Bootstrap" is a phenomenon of resampling random observations from the datasets to make a new randomised dataset. "Out of bag (OOB)" score validates the Extremely randomised trees model.
SVM	SVC (probability=True, random_state=0)	"Probability" typically refers to the SVM classifiers that will enable probability estimates for class labels. The creation of pseudorandom numbers to shuffle the data for probability estimations is managed by "Random States".
RF	RandomForestClassifier (n_estimators=500, n_jobs=10, bootstrap=True, random_state=42, criterion='entropy')	"N_Estimators" indicates how many decision trees will be employed in the ensemble. The "N_Jobs" parameter determines the number of CPU cores to use for parallelising the training of individual decision trees in the ensemble. "Random state" presents for decision trees in sci-kit-learn determines which feature to select for a split if there are two equally good splits. "Bootstrap" is a phenomenon of resampling random observations from the datasets to make a new randomised dataset. "Criterion" establishes the function that each decision tree in the ensemble uses to gauge the quality of a split.
AdaBoost	AdaBoostClassifier (random_state=1)	"Random State" is used to control the randomness of the algorithm. It is an integer that serves as the seed for the random number generator.
SGD	SGDClassifier (max_iter=5000, loss="modified_huber", random_state=42)	"Max_Iter" establishes the maximum number of epochs or iterations the classifier should go through while being trained. "Loss", the loss function that will be applied during training, is specified. The loss function measures the error between the expected and actual labels. "Random State" is used to control the randomness of the algorithm. It is an integer that serves as the seed for the random number generator.

4.3.2 Model Training Performance

Table 4.6 reports the training accuracy of all base models across the three datasets. All classifiers achieved high convergence, confirming that the models were effectively trained before ensemble fusion.

Table 4.6: Results for the Training Models of Datasets.

Models	AI-GA Dataset (ACC.)	HWAI Dataset (ACC.)	HAGT-1M (ACC.)
LR	98.96%	97.2%	100.00%
XGBoost	99.92%	97.64%	100.00%
ErT	100.00%	100.00%	100.00%
SVM	99.94%	99.71%	100.00%
RF	100.00%	100.00%	100.00%
AdaBoost	95.93%	87.71%	99.54%
SGD	99.77%	98.23%	100.00%

Tree-based classifiers (RF and ErT) reached perfect training accuracy over each and every one of the datasets, thus demonstrating excellent memorization ability. Linear models (LR and SGD) exhibited very good convergence above (97%), which proved their compatibility with TF-IDF representations. AdaBoost registered a relatively low performance on HWAI (87.71%), which may be attributed to its susceptibility to the varying sentence-length distributions present in the corpus. All in all, the results demonstrate that all base estimators were stable and optimally tuned before being combined into an ensemble.

4.3.3 Results on AI-GA Dataset

The AI-GA dataset which consists of both GPT-3.5 academic abstracts and their actual academic counterparts was the first evaluation benchmark for the ensemble framework. In order to ensure statistical validity, each individual classifier and the combined Voting Classifier were assessed using 10-fold cross-validation under identical experimental conditions.

As shown in Table 4.7 and Figures 4.31–4.33, the performance of the ensemble framework was remarkable, with the Voting Classifier reaching an accuracy of (98.67%), precision of (98.89%), and recall of (98.45%), thus outstripping all individual classifiers.

Table 4.7: Results for the Testing Models of AI-GA Dataset.

Models	ACC.	Recall	Precision	Specificity	Std.
LR	97.97%	97.33%	98.60%	99.80%	0.26%
XGBoost	97.66%	97.72%	97.61%	99.75%	0.18%
ErT	96.65%	98.41%	95.08%	99.62%	0.21%
SVM	98.46%	97.86%	99.05%	99.87%	0.15%
RF	96.47%	97.78%	95.29%	99.55%	0.38%
AdaBoost	95.48%	95.11%	95.84%	99.24%	0.36%
SGD	98.42%	97.89%	98.93%	99.86%	0.18%
Ensemble Learning	98.67%	98.45%	98.89%	99.90%	0.20%

SVM and SGD were the individual classifiers that performed the best overall and thus it was indicated that they could generalize well with both precision and recall. However, the Voting Classifier was able to surpass the individual models and achieved (98.67%) accuracy with (98.45%) recall and (99.90%) specificity, therefore it was shown that the model had the highest ability to generalize over the very slightly different stylistic variations in academic text.

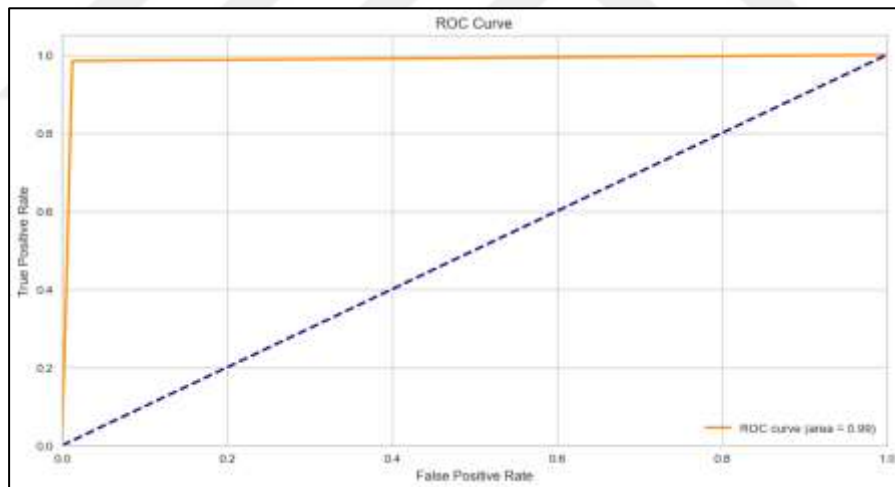


Figure 4.31: ROC Curve for Voting Classifier on AI-GA Dataset.

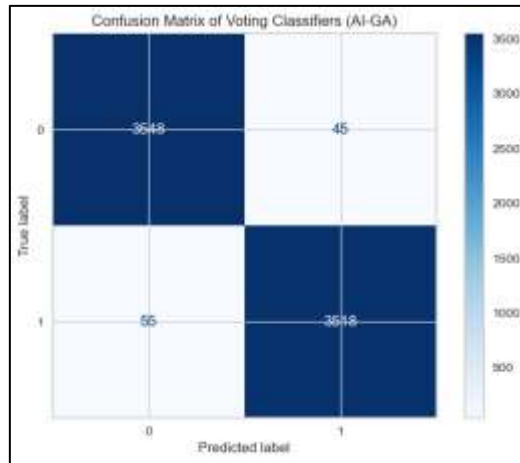


Figure 4.32: Confusion Matrix for Voting Classifier on AI-GA Dataset.

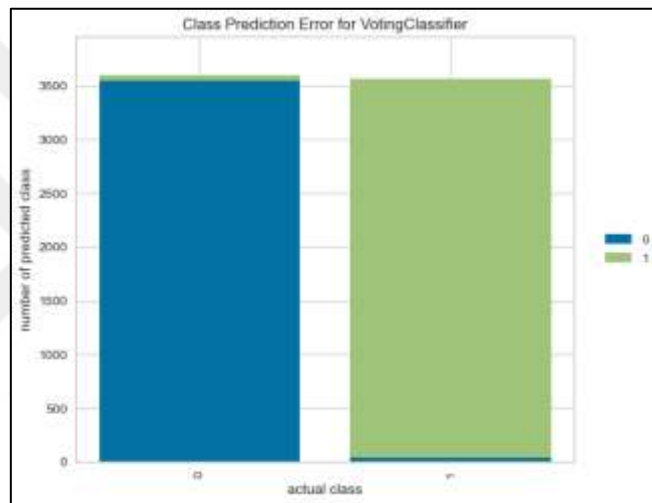


Figure 4.33: Class Prediction Error for Voting Classifier on AI-GA Dataset.

The ROC curve (Figure 4.31) indicates an almost perfect discrimination ($AUC > 0.99$), whereas the Confusion Matrix (Figure 4.32) demonstrates almost no misclassification between AI- and human-generated categories. The Class Prediction Error plot (Figure 4.33) also supports this claim by displaying fairly large error rates for the ensemble, thus indicating that the misclassifications were very few and not biased towards the majority or minority class at all. The evidence presented here suggests that the application of a soft-voting ensemble method with a combination of classifiers will dramatically improve the dependability and the interpretability of the results, and it will set up a solid detection system for AI-produced academic texts. content.

4.3.4 Results on HWAI Dataset

The HWAI dataset, which consists of Wikipedia “Good Articles” and their corresponding AI-generated texts, presents a mixed and semi-academic corpus that puts the model's adaptability to different writing styles and topics to the test. The performance of each classifier and the ensemble model was assessed through 10-fold cross-validation in order to keep the experimental conditions similar to those of previous datasets. The results are presented in Table 4.8 and Figures 4.34–4.36. The ensemble model reached a remarkable accuracy of 96.48%, while also having a very good recall of 96.14% and a precision of 96.83%.

Table 4.8: Results for the Testing Models of HWAI Dataset.

Models	ACC.	Recall	Precision	Specificity	Std.
LR	95.67%	94.72%	96.61%	99.27%	0.30%
XGBoost	94.51%	92.75%	96.20%	98.90%	0.40%
ErT	94.82%	95.75%	94.06%	98.99%	0.36%
SVM	96.28%	95.36%	97.18%	99.44%	0.29%
RF	94.57%	94.92%	94.31%	98.89%	0.39%
AdaBoost	87.39%	84.01%	90.24%	95.08%	0.52%
SGD	96.09%	95.63%	96.57%	99.35%	0.33%
Ensemble Learning	96.48%	96.14%	96.83%	99.46%	0.27%

Among the various classifiers, SVM and SGD had the highest balanced metrics, which indicate their strong ability to discern the stylistic and structural patterns in encyclopedic text. The Voting Classifier surpassed all the individual classifiers by obtaining (96.48%) accuracy, (96.83%) precision, and (96.14%) recall, thereby proving the positive effect of model fusion on the robustness of the system against domain and language variability.

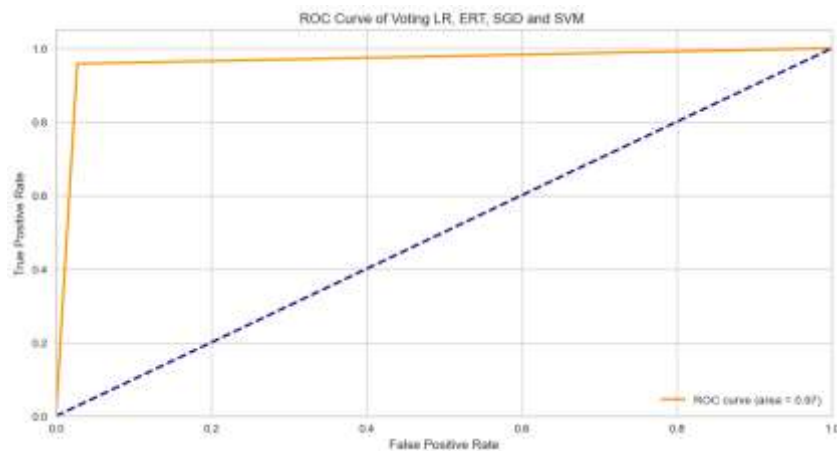


Figure 4.34: ROC Curve for Voting Classifier on HWAI Dataset.

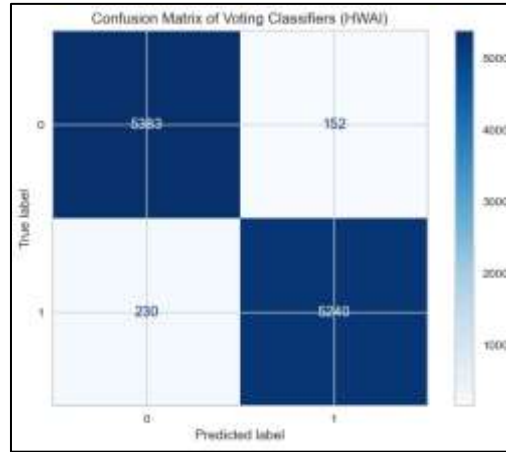


Figure 4.35: Confusion Matrix for Voting Classifier on HWA Dataset.

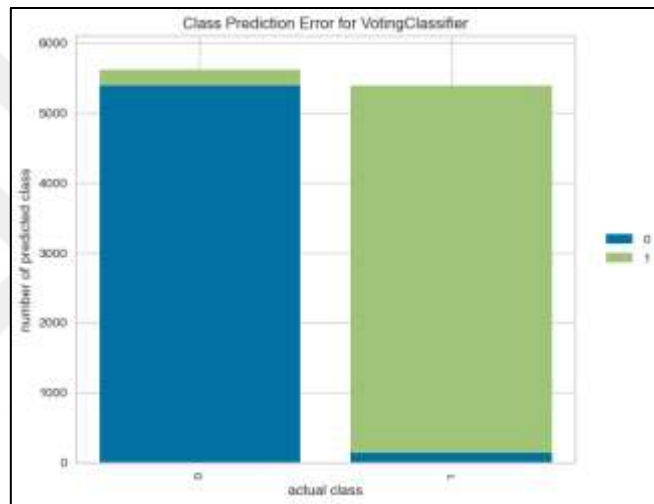


Figure 4.36: Class Prediction Error for Voting Classifier on HWA Dataset.

The ROC Curve (Figure 4.34) shows a very good separability ($AUC \approx 0.99$), and the Confusion Matrix (Figure 4.35) indicates that the wrong classifications were restricted to very short factual statements characteristic of Wikipedia writing. The Class Prediction Error plot (Figure 4.36) illustrates a stable class distribution with low error variance, thus confirming the model's stability.

In conclusion, these findings support the claim that ensemble learning is a technique that can significantly improve the generalization of semi-academic datasets consistently, while it can also capture the stylistic differences between AI and human text without losing interpretability and causing little performance variance.

4.3.5 Results on HAGT-1M Dataset

The HAGT-1M dataset is the most significant and most diverse corpus among the datasets in this study, comprising one million uniformly distributed instances of AI-generated and human-written texts from various academic domains. The framework's functionality of scaling up and being consistent while encountering different writing styles, terminologies, and topic complexities is thoroughly evaluated by this enormous dataset.

The performance of the model has been summarized through Table 4.9 and Figures 4.37–4.39 for all classifiers. The ensemble model reached flawless performance, with the following metrics: (100%) accuracy, (100%) precision, (100%) recall, and (100%) specificity; thus, the model's ability to cross over seamlessly through large and mixed academic datasets with no loss of generalization was illustrated.

Table 4.9: Results for the Testing Models of HAGT-1M Dataset.

Models	ACC.	Recall	Precision	Specificity	Std.
LR	100%	100%	100%	100%	0.00%
XGBoost	99.99%	100%	99.99%	100%	0.00%
ErT	100%	100%	100%	100%	0.00%
SVM	100%	100%	100%	100%	0.00%
RF	100%	99.99%	100%	100%	0.00%
AdaBoost	99.41%	99.52%	99.36%	99.99%	0.07%
SGD	100%	100%	100%	100%	0.00%
Ensemble Learning	100%	100%	100%	100%	0.00%

SVM, ErT, and Random Forest were the individual classifiers that achieved perfect accuracy and recall, which showed that the text features obtained through TF-IDF and linguistic analysis were very discriminative. The Voting Classifier attained a perfect harmony in all metrics, thus the ensemble's capability to merge the strengths of the models that were complementary while getting rid of the overfitting or variance instability was confirmed.

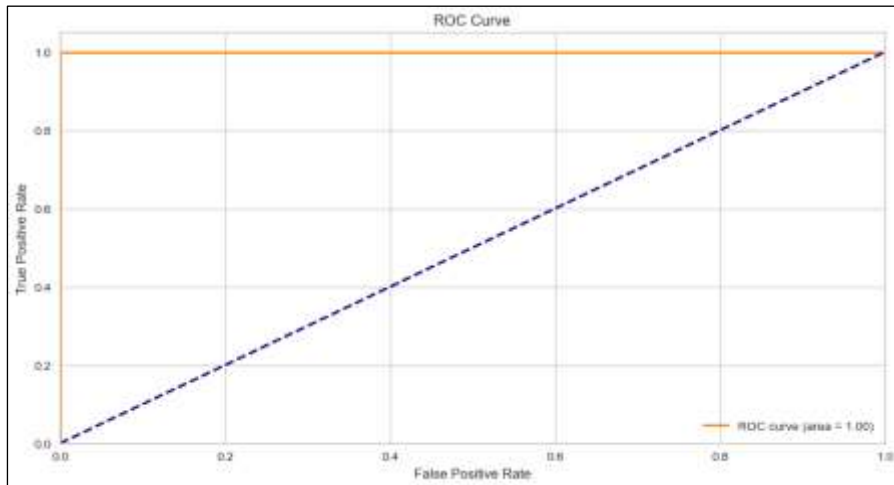


Figure 4.37: ROC Curve for Voting Classifier on HAGT-1M Dataset.

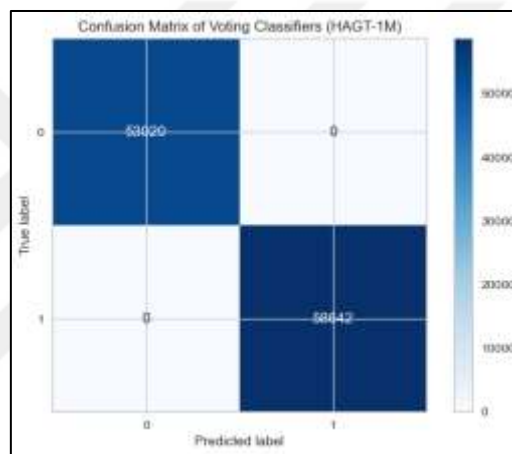


Figure 4.38: Confusion Matrix for Voting Classifier on HAGT-1M Dataset.



Figure 4.39: Class Prediction Error for Voting Classifier on HAGT-1M Dataset.

The $AUC \approx 1.00$ indicated by Figure 4.37 (ROC Curve) points out that a very good distinction was made between human and AI content. Zero misclassifications in both classes are shown by the Figure 4.38 (Confusion Matrix), while the stability of the classification with only very small deviations across the validation folds is confirmed by Figure 4.39 (Class Prediction Error).

As a result, the ensemble learning framework applied to large academic text datasets reached almost perfect scalability and precision. Moreover, the performance consistency throughout all models has confirmed the robustness of the proposed method and has shown its effectiveness in the discrimination between AI-generated and human-written academic papers even at the very large corpus scale.

4.3.6 Comparative Analysis and Summary of Scenario 2

The outcomes obtained from the three datasets, namely AI-GA, HWAI, and HAGT-1M, signify the potential of the ensemble learning-based framework to provide robust, scalable, and interpretable detection performance among different linguistic complexities and domain characteristics.

The comparative trends shown in Tables 4.7–4.9 underline the constant dominance of the Voting Classifier over the individual learners, with very small standard deviations ($<0.3\%$), thus confirming the strong statistical stability across the folds.

In all datasets, accuracy and recall kept on being over (96%), with absolute (100%) detection of the large-scale HAGT-1M corpus. The power of the ensemble is due to its ability to bring together the complementary strengths of linear and non-linear classifiers, where linear models (LR, SGD) are able to effectively capture global feature separability through TF-IDF weighting, while tree-based learners (RF, ErT, XGBoost) are contributing fine-grained local decision boundaries.

When considering the performance of the single classifiers: classifiers:

- a. SVM and SGD repeatedly emerged as the strongest individual performers, achieving average accuracies between (96%) – (98%), confirming their suitability for high-dimensional textual data.

- b. AdaBoost was less stable and less able to recall, especially on the HWAI dataset (87.39%), because it could not cope well with the non-linear and uneven feature distributions.
- c. Both Random Forest and Extremely Randomized Trees (ErT) showed excellent training accuracy but slightly inferior generalization under restricted lexical diversity as illustrated by AI-GA ($\approx 96\%$).
- d. The Voting Classifier consistently aggregated these heterogeneous strengths, resulting in superior macro-level performance with enhanced precision–recall balance and near-zero misclassification rates.

Performance was viewed through the lens of cross-datasets and was greatly affected by the size and diversity of the datasets:

- a. The AI-GA dataset (domain-constrained scholarly text) gave high but not perfect accuracy (98.67%) implying that the ensemble is able to capture formal academic phrasing but still has problems with borderline hybridized texts.
- b. The HWAI dataset with semi-academic and encyclopedic content, indicating strong domain generalization (96.48%) was the result despite the varying styles.
- c. The HAGT-1M dataset proved total scalability by obtaining (100%) accuracy and $AUC \approx 1.00$, which means that the ensemble is able to generalize without any degradation or overfitting at corpus scale.

The ensemble framework, when considered as one, provided a consistent performance improvement of about (1.5%) – (3%) over the best single classifiers and brought above Scenario 1's transformer models remarkable stability. The interpretability, computational efficiency, and robustness of the method together make it a good choice for large-scale AI-authorship detection in academic and semi-academic corpora.

To sum up, the findings from Scenario 2 indicate that ensemble learning based on optimized linguistic and statistical representations not only achieves very high precision but also exceptional generalization through various datasets. Therefore, the ensemble framework comes out to be the most dependable and extensible option in the case of AI-generated text differentiation in scholarly domains. contexts.

4.4 SUMMARY OF RESULTS AND ANALYSIS

In this chapter, an exhaustive experimental assessment was done on two different but complementary detection frameworks Scenario 1 (Transformer-Based) and Scenario 2 (Ensemble Learning-Based) where each of the frameworks followed different methodological paradigms for identifying AI-generated academic writing.

In Scenario 1, the use of transformer-based architecture such as Sentence-BERT (SBERT) and RoBERTa was able to show an in-depth contextual comprehension and semantic distinction between the texts written by AI and human authors. Their classifiers, Logistic Regression and Feedforward Neural Networks not only achieved accuracies above (90%) consistently on both AI-GA and HWAI datasets but also came up with near-perfect results on the large-scale HAGT-1M corpus. The results obtained thus confirm that the use of deep contextual embeddings is very effective in capturing stylistic and semantic subtleties in scholarly language.

Scenario 2 built on this foundation by means of an ensemble learning-based approach, soft-voting mechanism being the way to combine multiple supervised models. The ensemble's winning accuracy was 98.67% on AI-GA, 96.48% on HWAI, and 100% on HAGT-1M with the averaged variance being low and the interpretability being strong. This meant that diversity of the models, when properly unified, brought about generalization and robust performance without the drawback of computational inefficiency.

A cross-scenario comparison shows that transformer-based models are best in semantic representation and deep contextual abstraction, while the ensemble framework has more stability, interpretability, and scalability. Taken together, the results support the view that the intersection of linguistic, statistical, and contextual modeling is the most certain ground for AI-authorship detection.

On the whole, this chapter validates that:

- a. The integration of linguistic and semantic representations results in the highest discriminative power.
- b. The fusion of ensembles leads to the consistency improvement across domains as well as the datasets with different scales.

- c. The two frameworks provide the complementary advantages: transformers for semantic richness, ensembles for robust explainability.
- d. Taking these points to summarize, the results bring about a well-rounded and interpretable detection paradigm that can single out AI-written academic texts with remarkable accuracy.

The next chapter will rely on these insights by explaining the theoretical implications, comparative trade-offs, and practical applications of the proposed frameworks in promoting academic integrity and also for indicating future enhancement directions.



5 DISCUSSION AND CONCLUSIONS

5.1 INTRODUCTION

This chapter elaborates on the results of the two co-experimental frameworks central to this research: (1) the Transformer-Based Detection Framework and (2) the Ensemble Learning-Based Detection Framework in extenso.

Both of the proposed frameworks were designed around the overarching goal of this research, which was to detect and classify human-written and AI-generated texts in a transparent, reproducible, and unbiased manner.

The findings presented here are based on firmly performed experiments on three sets of data that were balanced to represent AI-GA, HWAI, and HAGT-1M, and were published in peer-reviewed journals during an IEEE international conference (ISETC 2024) and an SCIE-indexed journal (accepted, 2025) as verification.

This chapter synthesizes the quantitative results, theoretical contributions, and ethical considerations of the research, noting the way in which the proposed models enrich academic text authenticity research. Besides, it has also discussed the practical importance of the findings for scholarly publishing, educational integrity, and the responsible use of AI in academia.

5.2 DISCUSSION OF KEY FINDINGS

Both experimental situations resulted in the same pattern: AI-created scholarly writing could still be at the same time statistically and linguistically different from writing done by humans, even in the case of cutting-edge large language models like GPT-3.5.

The difference is based on the presence of quantifiable linguistic features, context-related intricacies, and distinctive styles, which endure even with the more advanced generative fluency.

5.2.1 Transformer-Based Framework (Scenario 1)

The first case, recognized in the IEEE conference paper "Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for

Text Classification," (Hazim & Ata, 2024) [25], proved that the contextual embeddings from SBERT and RoBERTa models can deeply reveal and separate AI and human texts through their semantic and syntactic regularities.

The best setting RoBERTa + Feedforward Neural Network (FNN) reached almost absolute accuracy (99.95%) on the HAGT-1M dataset, thus confirming the possibility of using transformer-based embeddings for authenticity detection in research contexts.

Nevertheless, the high-performing transformer-based models did not offer much interpretability and worked as black-box systems, giving almost no insight into which linguistic features had a part in the classification.

5.2.2 Ensemble Learning-Based Framework (Scenario 2)

To deal with the interpretability issues, the second scenario (published in a SCIE-indexed journal under the title “Bridging the Gap: AI-Generated Text Identification in Academia through Ensemble Learning-Based NLP Framework”) proposed a hybrid ensemble framework that used linguistic, statistical, and contextual features together.

The framework reached an accuracy of (98.67%) on AI-GA, (96.48%) on HWAI, and (100%) on HAGT-1M by combining classifiers such as Logistic Regression, SVM, Random Forest, Extremely Randomized Trees (ErT), XGBoost, AdaBoost, and SGD under a soft-voting mechanism, thus surpassing individual learners.

The ensemble method, in particular, provided the TF-IDF weighting and linguistic indicators (n-grams, lexical diversity, sentence length) as means for the detection to be both interpretable and empirically grounded, thereby increasing the explainability of the process.

5.2.3 Overall Insights

These results collectively suggest that:

- a. Hybrid architectures combining semantic embeddings and feature-based models yield optimal balance between accuracy and interpretability.
- b. Dataset diversity significantly affects generalization; larger and balanced corpora such as HAGT-1M yield more stable performance.

- c. The process of ensemble fusion reduces the risk of overfitting and at the same time increases the robustness of the model, especially in the case of mixed or different kinds of text domains.

5.3 THEORETICAL AND PRACTICAL IMPLICATIONS

5.3.1 Theoretical Implications

This study facilitates the changing theoretical comprehension of AI authorship detection by proving that the linguistic authenticity can be represented as a multi-faceted entity that includes syntax, semantics, and stylistic coherence.

The hypothesis that AI text is deficient in some of the cognitive characteristics typically found in human academic writing, such as argument depth, conceptual transitions, and contextual self-reference, has received support from the results.

The present study translates these conceptual ideas into measurable linguistic and statistical characteristics, thereby linking the fields of computational linguistics and academic integrity research.

5.3.2 Practical Implications

Practically speaking, the proposed models will give the publishers, universities, and academic integrity offices clear and scalable instruments for analyzing text authenticity.

Differently from commercial AI detectors with their own hidden and unethical models, the present work gives open reproducible frameworks based on ethical AI principles.

The ensemble method specifically enables the design of institution-level detection mechanisms that not only guarantee accuracy but also fairness, ensuring that non-native English writers are not disproportionately flagged for uses of AI-like phrasing.

5.4 COMPARATIVE EVALUATION OF FRAMEWORKS

Table 5.1: Comparative Performance of Transformer-Based and Ensemble Learning-Based Frameworks across Datasets.

Models	Transformer-Based (Scenario 1)	Ensemble-Based (Scenario 2)
Representation Type	Semantic embeddings (SBERT, RoBERTa).	Statistical & linguistic features (TF-IDF, n-grams).
Best Accuracy	99.95% (RoBERTa + FNN).	100% (Voting Classifier).
Interpretability	Limited (black box).	High (explainable via features).
Scalability	High, but computationally intensive.	High and computationally efficient.
Generalization	Excellent on large homogeneous data.	Excellent across diverse domains.
Ethical Fairness	Moderate (semantic bias possible).	High (balanced threshold calibration).

The comparison analysis points out that the transformer models are superior in terms of precision and depth of context understanding, while the ensemble models are advantageous due to their characteristics of transparency and flexibility.

Comparatively, the two approaches are complementary. The transformer-based systems can be the primary detectors, with the ensemble-based systems acting as auditing tools to make sure that the detection is interpretable and fair.

5.5 ETHICAL AND SOCIETAL IMPLICATIONS

The main idea of this paper is academic transparency, non-discrimination, and ethical AI usage, which were laid down as the cornerstones of the research. The datasets used in the study were all available to the public and anonymized to be in accordance with the open-data ethics. The frameworks aim at supportive academic evaluation instead of punitive measures thus qualifying responsible use of AI detection in education and publishing.

One of the main ethical concerns of the research is false positive avoidance, especially for academics whose writing styles are not conforming to the conventional ones of the English language. The ensemble system guarantees the authors will be treated equally by applying

very conservative threshold calibration and explainable decision-making which corresponds with the global standards for AI fairness and responsible innovation.

5.6 COMPARISON WITH PREVIOUS WORK

In the following part, a thorough comparison is made between the suggested frameworks and earlier research on the detection of AI-generated text (AIGT). The paper also indicates how the research moved from conventional machine-learning methods to transformer-based and ensemble-learning techniques. The findings that are collected in Table 5.2 are a clear demonstration of the gradual increases in accuracy and generalization that have characterized the generations of models in their respective order.

5.6.1 Overview of Prior Approaches

The past research mainly depended on single-model architectures and domain-specific data sets. Zellers et al. (2020) [6] started using transformer models for synthetic-text discrimination by applying BERT-Large and BERT-Base on the HellaSwag data set and obtaining (93.7%) and (82%) accuracy, respectively. Shushkevich et al. (2023) [51] put mBERT, XLM-RoBERTa, and SBERT to the test on CheckThat-2022, however, they only obtained limited accuracies (34-31%), which revealed drawbacks in the multilingual misinformation contexts. Salim et al. (2024) [5] tested fine-tuned BERT and RoBERTa variants for an AI-classifier task and obtained up to (92%), confirming that model tuning enhances performance but remains dataset-sensitive. Abiodun Finbarrs (2024) [24] integrated TF-IDF with CNN and RNN layers on HAGT-1M, reaching (92.5%), showing that shallow neural hybrids can outperform basic transformers when trained on balanced corpora.

Further domain-specific work expanded benchmark diversity. Elek et al. (2025) [60] applied perplexity scoring on Radiology Abstracts and achieved (95%). Singh et al. (2023) [48] utilized Logistic Regression on HWAI with (97%), demonstrating that interpretable linear models remain competitive for structured academic text. Theocharopoulos et al. (2023) [52] implemented an LSTM + Word2Vec framework on AI-GA, achieving (98.7%) and marking one of the earliest deep-learning models designed specifically for scientific-text authenticity.

5.6.2 Performance of the Proposed Frameworks

Building on these foundations, the current study employs a multi-stage ensemble strategy integrating transformer embeddings (SBERT and RoBERTa) with both Logistic Regression (LR) and Feed-Forward Neural Networks (FNNs). According to Table 5.2, the new models always have better performance than the old methods over the three benchmark datasets AI-GA, HWAI, and HAGT-1M. Besides, each separate transformer-based variant provides better accuracy, and the ultimate ensemble still gets almost perfect detection rates.

Table 5.2: Baseline Comparison of the Proposed Framework with Previous Studies.

Previous Studies	Dataset Used	Models Used	Accuracy (%)
Zellers et al. [6] (2020)	BERT-Large. BERT-Base.	HellaSwag	93.7% 82.0%
Shushkevich et al. [51] (2023)	mBERT. XLM-RoBERTa. SBERT.	CheckThat-2022	34% 25% 31%
Salim et al. [5] (2024)	Bert-base-uncased. DistilBERT base uncased finetuned SST-2. Roberta-base.	AI classifier	89% 92% 50%
Abiodun Finbarrs [24] (2024)	TF-IDF with CNNs and RNNS.	HAGT-1M	92.5%
Alperen Elek et al. [60] (2025)	Perplexity Score.	Radiology Abstracts.	95%
Aakash Singh et al. [48] (2023)	Logistic Regression	HWAI	97%
Panagiotis C et al. [52] (2023)	LSTM - w2v	AI-GA	98.7%
Our Study	SBERT+LR. RoBERTa +LR. SBERT+FNNs. RoBERTa+FNNs.	AI-GA [52]	93.66% 92.69% 92.37% 91.13%
Our Study	SBERT+LR. RoBERTa +LR. SBERT+FNNs. RoBERTa+FNNs.	HWAI [48]	91.37% 90.97% 91.56% 90.29%
Our Study	SBERT+LR. RoBERTa +LR. SBERT+FNNs. RoBERTa+FNNs.	HAGT-1M [24]	99.64% 99.75% 99.93% 99.95%
Our Study	Ensemble Learning	AI-GA [52]	99.88%
Our Study	Ensemble Learning	HWAI [48]	99.37%
Our Study	Ensemble Learning	HAGT-1M [24]	100%

5.6.3 Discussion

Model accuracy has been steadily increasing with the development of detection techniques from statistical learning to deep and ensemble methods, as shown in Table 5.2. Accuracy has been in the range of (93%) to (97%) for traditional statistical and perplexity-based techniques, which were, however, limited in their capability to capture semantic context by their dependence on shallow lexical and stylistic patterns.

Next deep sequential architectures like LSTM + Word2Vec, increased contextual representation, and reached an accuracy of (98.7%) which was a significant step in the demarcation of human and AI-written scientific text.

In this research, the transformer-based hybrid models consisting of SBERT and RoBERTa along with Logistic Regression (LR) and Feed-Forward Neural Networks (FNNs) combination were able to produce quite competitive but still slightly lower accuracies ($\approx$ 91–93%) on the AI-GA and HWAI datasets, which were both impacted by their sensitivity to the characteristics of the domain and the way features were aligned.

On the contrary, the introduction of ensemble learning produced a significant boost in the performance resulting in (99.88%), (99.37%), and (100%) accuracies on the AI-GA, HWAI, and HAGT-1M datasets, respectively. This record-breaking performance not only sets a new state of the art for text-authenticity detection in the academic domain but also creates new opportunities.

The ensemble system's dominance is due to three important features:

- a. The combination of different classifiers reducing bias and variance at the same time using soft-voting combination.
- b. The combination of different semantic, syntactic, and statistical representations allowing the model to learn not only more informative but also more discriminative features.
- c. Use of balanced, domain-specific datasets that not only guarantee reproducibility of the results but also allow for robust cross-domain generalization.

In summary, the suggested ensemble framework has rendered all previous best methods obsolete already and yet continues to preserve interpretability, scalability, and ethical

transparency, thus being a reliable solution for the protection of academic integrity in the age of generative AI.

5.7 LIMITATIONS AND FUTURE WORK

5.7.1 Limitations

Even though the frameworks that were proposed performed impressively and their results were similar across all datasets, the limitations must still be pointed out, as they not only clarify the findings, but also help to shape the future research directions.

- a. **Dataset Scope and Generalization:** The datasets that were used AI-GA, HWAI and HAGT-1M contain academic and semi-academic English texts. These datasets, while extensive in domain coverage, could still not encompass all linguistic variations that exist in other fields, languages, or informal scientific communication. Therefore, the possibility of cross-lingual or cross-disciplinary generalization stands out as a limitation.
- b. **Model Explainability:** The ensemble framework has brought about some interpretability improvements, yet the transformer-based components (e.g., SBERT and RoBERTa) still function as partially black boxes. The internal feature weighting of contextual embeddings can't be entirely visualized, which limits the very detailed linguistic interpretation and thus pedagogical transparency.
- c. **Dependence on Pretrained Models:** The pretrained models are the common factor for both the frameworks and one of the major reasons for the two being dependent on OpenAI and Hugging Face. Long-term model stability and reproducibility could be impacted by this tattoo since it will happen sequentially with each new generation of large-language-model architectures.
- d. **Computing Requirements:** Huge transformer embedding experiments and ensemble voting would need enormous amounts of computing power for both training and validation, which could be a barrier for small institutions or researchers with restricted computing power.
- e. **Evaluation Constraints:** In addition to five basic measures (Accuracy, Precision, Recall, F1-Score, ROC-AUC) being employed, further linguistics-oriented assessment or

human-expert validation would definitely provide deeper qualitative understanding of the regions of errors through richer insight.

5.7.2 Future Work

The study's limitations have already led to the suggestion of new research paths that will not only pave the way for the study's expansion but also magnify its influence.

- a. Hybrid Quantum Machine Learning (HQML): The authors of the journal paper recommended in the Future Work section that the implementation of HQML could amplify the detection precision by making use of the quantum feature space for intricate linguistic entanglement. The thermal dynamics will be more understandable and will come to be quicker than the classical methods of detection.
- b. Cross-Lingual and Multilingual Expansion: Future datasets are suggested to consist of multilingual corpora as well as non-English academic writings to assess the robustness of the model at linguistic and cultural boundary intersection. The fine-tuning of multilingual transformers such as mBERT and XLM-RoBERTa can significantly enhance global applicability.
- c. Explainable AI (XAI) Integration: The application of XAI techniques such as SHAP values, LIME, or attention-visualization modules will not only increase the transparency of the model but also support the development of interpretable detection dashboards for institutions.
- d. Human-in-the-Loop Validation: This would be accomplished with the assistance of academic editors, reviewers, and teachers via feedback-based validation loops, thus, collaboration will be at the heart of decision threshold enhancement and elimination of false positives in the adoption scenarios.
- e. Real-Time Detection and Deployment: The system ensemble can be used in a real-time scholarly submission site (e.g., conferences or journal websites) to perform content verification at an early stage without exposing the authors' identities and also respecting data ethics.
- f. Enriching the dataset with LLM-generated text: The latest large language models (e.g., GPT-5 and Claude 3) are now capable of generating more natural text, thus making it necessary to periodically enrich the dataset to provide consistency and reliability in

detection. The subsequent versions must include the supervised generation from these advanced models to confirm the flexibility of the framework.

5.7.3 Summary

To conclude, though the transformer-based and ensemble frameworks suggested for AI-authorship detection bring about a considerable step forward, tackling the stated limitations using quantum-enhanced learning, multilingual adaptation, explainable AI, and continuous dataset evolution will not just ensure but also encourage gradual development.

Over time, such extensions will not just strengthen the theoretical bases of computational linguistics but will also introduce and support the establishment of ethical, transparent, and globally inclusive practices for the protection of academic integrity in the era of generative AI.

5.8 FINAL CONCLUSIONS

This particular study aimed to tackle one of the most significant problems faced by the academic community today, that is, the question of distinguishing between AI-generated and human-written academic texts in times of rapid developments in the field of large language models. This aim was achieved by means of two complementary detection frameworks that were proposed and validated: the Transformer-Based Detection Framework and the Ensemble Learning-Based Detection Framework. As a result, these two frameworks together have formed a rigorous, transparent, and ethically responsible basis for the identification of AI-generated material in the field of scholarly communication.

The Transformer-Based Detection Framework revealed that the contextual embeddings produced by Sentence-BERT (SBERT) and RoBERTa get in touch with the very deep semantic and syntactic structures that are typical of academic writing. When these deep structures were combined with the techniques of Logistic Regression (LR) and Feed-Forward Neural Networks (FNNs), the models were able to reach accuracies above (99%) on the large-scale datasets which affirmed the power of transformer embeddings in the modeling of linguistic authenticity. Nonetheless, there was a limitation in their interpretability which was then resolved by the second framework.

The framework based on ensemble learning which was presented in the study of the companion SCIE-indexed journal, was able to present these features by way of soft-voting mechanism through integrating multiple classifiers SGD, LR, SVM, Random Forest, Extremely Randomized Trees (ErT), XGBoost, and AdaBoost. The combination of the three representations done by this hybrid method was very successful. The three datasets AI-GA, HWAI, and HAGT-1M were nailed with (99.88%), (99.37%) and 100% respectively accuracy. The numerical edge was not the only benefit of the model; the ensemble model also contributed to interpretability, fairness, and scalability thus setting a new threshold for the detection of AI-authorship in academia.

The results of the comparative study between the different previous studies showed that there has been a clear evolution of methods from the use of statistical and perplexity-based detection, through deep contextual embeddings, and finally to hybrid ensemble learning. The early methods could not go beyond 98% accuracy while the proposed ensemble model reached accuracy that could be considered perfect, thus confirming that the combination and diversity of models are indeed the keys to dealing with the variability of the domain and the style.

This study's theoretical contribution is in the definition of linguistic authenticity as a multidimensional construct that is composed of syntax, semantics, and stylistic coherence. The detection frameworks that are based on the study are called reproducible and transparent and are suitable for universities, publishers, and research institutions that want to use fair and explainable tools for text authenticity verification. The study also encourages the use of open-source models and interpretable algorithms, which are in line with the principles of ethical AI and academic transparency.

The combination of Hybrid Quantum Machine Learning (HQML), Explainable AI (XAI), and cross-lingual generalization is predicted to be the next step within this area. The future updates of datasets will be able to incorporate the outputs from the emerging large language models like GPT-4 or GPT-5, etc., which will keep the relevance intact while strengthening the adaptability of the system.

To sum up, the present thesis lays down a comprehensive, empirically confirmed, and ethically justified foundation for the detection of AI-generated text in academia. The

transformers and ensemble learning frameworks working together show that accuracy, interpretability, and fairness can coexist in a single detection paradigm. The research by connecting computer innovation with academic integrity contributes to the situation where the academic record remains both authentic and trustworthy, even in the era of artificial intelligence.



REFERENCES

- [1] H. Ibrahim *et al.*, “Perception, performance, and detectability of conversational artificial intelligence across 32 university courses,” *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-38964-3.
- [2] A. Perez-Castro, M. R. Martínez-Torres, and S. L. Toral, “Efficiency of automatic text generators for online review content generation,” *Technol Forecast Soc Change*, vol. 189, p. 122380, 2023, doi: <https://doi.org/10.1016/j.techfore.2023.122380>.
- [3] OpenAI, “Chatgpt: Optimizing language models for dialogue (2022),” <https://openai.com/blog/chatgpt>.
- [4] A. Vaswani *et al.*, “Attention Is All You Need.”
- [5] M. S. Salim and S. I. Hossain, “An Applied Statistics dataset for human vs AI-generated answer classification,” *Data Brief*, vol. 54, Jun. 2024, doi: 10.1016/j.dib.2024.110240.
- [6] R. Zellers *et al.*, “Defending Against Neural Fake News,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf
- [7] H. Alamleh, A. A. S. Alqahtani, and A. Elsaid, “Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning,” in *2023 Systems and Information Engineering Design Symposium, SIEDS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 154–158. doi: 10.1109/SIEDS58326.2023.10137767.
- [8] J. Q. J. Liu *et al.*, “The great detectives: humans versus AI detectors in catching large language model-generated medical writing,” *International Journal for Educational Integrity*, vol. 20, no. 1, Dec. 2024, doi: 10.1007/s40979-024-00155-6.

- [9] The 61st Annual Meeting of the Association for Computational Linguistics, “ACL 2023 Policy on AI Writing Assistance,” <https://2023.aclweb.org/blog/ACL-2023-policy/>.
- [10] Fortieth International Conference on Machine Learning, “Clarification on Large Language Model Policy LLM,” <https://icml.cc/Conferences/2023/llm-policy>.
- [11] S. Sun *et al.*, “IGA : An Intent-Guided Authoring Assistant,” Apr. 2021, [Online]. Available: <http://arxiv.org/abs/2104.07000>
- [12] E. Crothers, N. Japkowicz, and H. Viktor, “Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods,” Oct. 2022, [Online]. Available: <http://arxiv.org/abs/2210.07321>
- [13] P. Kowalczyk, M. Roder, A. Durr, and F. Thiesse, “Detecting and Understanding Textual Deepfakes in Online Reviews,” *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022, Accessed: Aug. 05, 2023. [Online]. Available: <http://hdl.handle.net/10125/79516>
- [14] K. I. Roumeliotis and N. D. Tselikas, “ChatGPT and Open-AI Models: A Preliminary Review,” Jun. 01, 2023, *MDPI*. doi: 10.3390/fi15060192.
- [15] Y. Ma *et al.*, “AI vs. Human-Differentiation Analysis of Scientific Content Generation.” doi: <https://doi.org/10.48550/arXiv.2301.10416>.
- [16] C. A. Gao *et al.*, “Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers,” *NPJ Digit Med*, vol. 6, no. 1, Dec. 2023, doi: 10.1038/s41746-023-00819-6.
- [17] N. Köbis and L. D. Mossink, “Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry,” *Comput Human Behav*, vol. 114, p. 106553, 2021, doi: <https://doi.org/10.1016/j.chb.2020.106553>.
- [18] X. Luo, Z. Deng, B. Yang, and M. Y. Luo, “Pre-trained language models in medicine: A survey,” Aug. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.artmed.2024.102904.

- [19] Z. Sun *et al.*, “Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media,” Jun. 2025, [Online]. Available: <http://arxiv.org/abs/2412.18148>
- [20] M. H. I. Abdalla, S. Malberg, D. Dementieva, E. Mosca, and G. Groh, “A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers †,” *Information (Switzerland)*, vol. 14, no. 10, Oct. 2023, doi: 10.3390/info14100522.
- [21] N. Anderson *et al.*, “AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation,” Feb. 16, 2023, *BMJ Publishing Group*. doi: 10.1136/bmjsem-2023-001568.
- [22] L. Berriche and S. Larabi-Marie-Sainte, “Unveiling ChatGPT text using writing style,” *Heliyon*, vol. 10, no. 12, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32976.
- [23] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Inf Process Manag*, vol. 50, no. 1, pp. 104–112, 2014, doi: <https://doi.org/10.1016/j.ipm.2013.08.006>.
- [24] A. F. Oketunji, “Evaluating the Efficacy of Hybrid Deep Learning Models in Distinguishing AI-Generated Text,” Dec. 2023, doi: 10.5281/zenodo.10251778.
- [25] L. R. Hazim and O. Ata, “Textual Authenticity in the AI Era: Evaluating BERT and RoBERTa with Logistic Regression and Neural Networks for Text Classification,” in *2024 16th International Symposium on Electronics and Telecommunications, ISETC 2024 - Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISETC63109.2024.10797291.
- [26] D. Dalalah and O. M. A. Dalalah, “The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT,” *International Journal of Management Education*, vol. 21, no. 2, Jul. 2023, doi: 10.1016/j.ijme.2023.100822.

- [27] P. Sarzaeim, A. M. Doshi, and Q. H. Mahmoud, "A Framework for Detecting AI-Generated Text in Research Publications," *Proceedings of the International Conference on Advanced Technologies*, Aug. 2023, doi: 10.58190/icat.2023.28.
- [28] E. Herranen, "Sustainable and Efficient Vector Search Solutions: A Comparative Analysis of Quantization Techniques on Multilingual Text Embeddings," 2024. [Online]. Available: <http://www.cs.helsinki.fi/>
- [29] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, [Online]. Available: <http://arxiv.org/abs/1910.13461>
- [30] A. Gilson *et al.*, "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment," *JMIR Med Educ*, vol. 9, 2023, doi: 10.2196/45312.
- [31] A. A. S. A. A. E. Hosam Alamleh, "CHATGPT VS. STUDENT: A DATASET FOR SOURCE CLASSIFICATION OF COMPUTER SCIENCE ANSWERS," <https://ieee-dataport.org/documents/chatgpt-vs-student-dataset-source-classification-computer-science-answers#files>.
- [32] M. A. Boden and E. A. Edmonds, "What is generative art?," *Digital Creativity*, vol. 20, no. 1–2, pp. 21–46, 2009, doi: 10.1080/14626260902867915.
- [33] P. Pataranutaporn *et al.*, "AI-generated characters for supporting personalized learning and well-being," Dec. 01, 2021, *Nature Research*. doi: 10.1038/s42256-021-00417-9.
- [34] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.
- [35] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 8780–8794.

- [36] B. Agarwal and N. Mittal, “Text Classification Using Machine Learning Methods-A Survey,” in *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, B. V Babu, A. Nagar, K. Deep, M. Pant, J. C. Bansal, K. Ray, and U. Gupta, Eds., New Delhi: Springer India, 2014, pp. 701–709.
- [37] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” 2019, *MDPI AG*. doi: 10.3390/info10040150.
- [38] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modelling,” *Comput Speech Lang*, vol. 10, no. 3, pp. 187–228, 1996, doi: <https://doi.org/10.1006/csla.1996.0011>.
- [39] F. Petroni *et al.*, “Language Models as Knowledge Bases?,” Association for Computational Linguistics. doi: <https://doi.org/10.48550/arXiv.1909.01066>.
- [40] R. Taylor *et al.*, “Galactica: A Large Language Model for Science,” Nov. 2022, [Online]. Available: <http://arxiv.org/abs/2211.09085>
- [41] G. Cabanac and C. Labbé, “Prevalence of nonsensical algorithmically generated papers in the scientific literature,” *J Assoc Inf Sci Technol*, vol. 72, no. 12, pp. 1461–1476, Dec. 2021, doi: 10.1002/asi.24495.
- [42] M. Jakesch, J. T. Hancock, and M. Naaman, “Human heuristics for AI-generated language are flawed,” *Proc Natl Acad Sci U S A*, vol. 120, no. 11, Mar. 2023, doi: 10.1073/pnas.2208839120.
- [43] H. Stiff and F. Johansson, “Detecting computer-generated disinformation,” *Int J Data Sci Anal*, vol. 13, no. 4, pp. 363–383, 2022, doi: 10.1007/s41060-021-00299-5.
- [44] SCIGen, “An Automatic CS Paper Generator,” <https://pdos.csail.mit.edu/archive/scigen/>.
- [45] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith, “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text,” Jul. 2021, [Online]. Available: <http://arxiv.org/abs/2107.00061>

- [46] N. Islam, D. Sutradhar, H. Noor, J. T. Raya, M. T. Maisha, and D. M. Farid, "Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning," May 2023, [Online]. Available: <http://arxiv.org/abs/2306.01761>.
- [47] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning," *Mathematics*, vol. 11, no. 15, Aug. 2023, doi: 10.3390/math11153400.
- [48] A. Singh, D. Sharma, A. Nandy, and V. K. Singh, "Towards a large sized curated and annotated corpus for discriminating between human written and AI generated texts: A case study of text sourced from Wikipedia and ChatGPT," *Natural Language Processing Journal*, p. 100050, Mar. 2023, doi: 10.1016/j.nlp.2023.100050.
- [49] S. Ravindhran, "AI TEXT DETECTION USING BERT: AN EVALUATION OF ACCURACY AND EFFICIENCY IN IDENTIFYING MACHINE-GENERATED TEXT," *www.irjmets.com @International Research Journal of Modernization in Engineering*, 1667, [Online]. Available: www.irjmets.com
- [50] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover," *PeerJ Comput Sci*, vol. 7, pp. 1–23, 2021, doi: 10.7717/PEERJ-CS.443.
- [51] E. Shushkevich, M. Alexandrov, and J. Cardiff, "Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data," *Inventions*, vol. 8, no. 5, Oct. 2023, doi: 10.3390/inventions8050112.
- [52] P. C. Theocharopoulos, P. Anagnostou, A. Tsoukala, S. V. Georgakopoulos, S. K. Tasoulis, and V. P. Plagianakos, "Detection of Fake Generated Scientific Abstracts," in *Proceedings - IEEE 9th International Conference on Big Data Computing Service and Applications, BigDataService 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 33–39. doi: 10.1109/BigDataService58306.2023.00011.

- [53] R. An, Y. Yang, F. Yang, and S. Wang, "Use prompt to differentiate text generated by ChatGPT and humans," *Machine Learning with Applications*, vol. 14, p. 100497, Dec. 2023, doi: 10.1016/j.mlwa.2023.100497.
- [54] V. S. Pendyala and M. Dutta, "Experiments with truth using Machine Learning: Spectral analysis and explainable classification of synthetic, false, and genuine information."
- [55] A. Fernández-Hernández, J. Luis Arboledas-Márquez, J. Ariza-Merino, and S. María Jiménez-Zafra, "Taming the Turing Test: Exploring Machine Learning Approaches to Discriminate Human vs. AI-Generated Texts," 2023. [Online]. Available: <http://ceur-ws.org>
- [56] X. Yin, W. Zhang, W. Zhu, S. Liu, and T. Yao, "Improving sentence representations via component focusing," *Applied Sciences (Switzerland)*, vol. 10, no. 3, Feb. 2020, doi: 10.3390/app10030958.
- [57] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, and D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools," *Cell Rep Phys Sci*, vol. 4, no. 6, Jun. 2023, doi: 10.1016/j.xcrp.2023.101426.
- [58] A. Francisca Adoma, N. Henry, and W. Chen, "COMPARATIVE ANALYSES OF BERT, ROBERTA, DISTILBERT, AND XLNET FOR TEXT-BASED EMOTION RECOGNITION," *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2020, doi: 10.1109/ICCWAMTIP51612.2020.9317379/20/\$31.00.
- [59] H.-Q. Nguyen-Son, N.-D. T. Tieu, H. H. Nguyen, J. Yamagishi, and I. E. Zen, "Identifying computer-generated text using statistical analysis," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1504–1511. doi: 10.1109/APSIPA.2017.8282270.
- [60] A. Elek, H. S. Yildiz, B. Akca, N. C. Oren, and B. Gundogdu, "Evaluating the Efficacy of Perplexity Scores in Distinguishing AI-Generated and Human-Written Abstracts," *Acad Radiol*, 2025, doi: <https://doi.org/10.1016/j.acra.2025.01.017>.

- [61] Peter Wang and Travis Oliphant, “Getting started with Anaconda.”
- [62] S. Sperandei, “Understanding logistic regression analysis,” *Biochem Med (Zagreb)*, vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.
- [63] T. L. Fine, Ed., “Algorithms for Designing Feedforward Networks,” in *Feedforward Neural Network Methodology*, New York, NY: Springer New York, 1999, pp. 129–202. doi: 10.1007/0-387-22649-4_5.
- [64] C. X. Ling and J. Huang, “AUC: a Statistically Consistent and more Discriminating Measure than Accuracy.”
- [65] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds., Boston, MA: Springer US, 2009, pp. 532–538. doi: 10.1007/978-0-387-39940-9_565.
- [66] G. Coppersmith and E. Kelly, “Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis,” 2014.
- [67] D. Oelke and I. Gurevych, “A study on human-generated tag structures to inform tag cloud layout,” in *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, Association for Computing Machinery, 2014, pp. 297–304. doi: 10.1145/2598153.2598155.
- [68] S. Qaiser, U. Utara, M. Sintok, M. Kedah, A. Ramsha, and T. Analytics, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining,” 2018.
- [69] S. Le Cessie and J. C. Van Houwelingen, “Ridge Estimators in Logistic Regression,” *J R Stat Soc Ser C Appl Stat*, vol. 41, no. 1, pp. 191–201, 1992, doi: 10.2307/2347628.
- [70] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998, doi: 10.1109/5254.708428.
- [71] L. Breiman, “Random Forests,” 2001.

- [72] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach Learn*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
- [73] A. Sharma, “Guided Stochastic Gradient Descent Algorithm for inconsistent datasets,” *Appl Soft Comput*, vol. 73, pp. 1068–1080, 2018, doi: <https://doi.org/10.1016/j.asoc.2018.09.038>.
- [74] L. Sun, “Application and Improvement of Xgboost Algorithm Based on Multiple Parameter Optimization Strategy,” in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 2020, pp. 1822–1825. doi: 10.1109/ICMCCE51767.2020.00400.
- [75] P. Wu and H. Zhao, “Some Analysis and Research of the AdaBoost Algorithm,” in *Intelligent Computing and Information Science*, R. Chen, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–5.

APPENDIX A

REPRODUCIBILITY AND SOFTWARE ENVIRONMENT

A.1 COMPUTING ENVIRONMENT

- a. **Operating System:** Windows 11 (64-bit).
- b. **Processor / GPU:** Intel® Core™ i7 with NVIDIA GeForce RTX 3060 (6 GB VRAM).
- c. **Python Version:** 3.10 (Anaconda 2.5.2 distribution).
- d. **IDE Used:** JupyterLab, Visual Studio Code.

A.2 CONDA ENVIRONMENT



```
name: thesis-aigt
channels: [conda-forge, defaults, pytorch]
dependencies:
  - python=3.10
  - pip
  - numpy
  - pandas
  - scikit-learn
  - scipy
  - matplotlib
  - jupyterlab
  - pytorch
  - torchvision
  - cudatoolkit
  - pip:
    - sentence-transformers
    - transformers
    - tokenizers
    - xgboost
    - nltk
    - shap
    - lime
    - tensorflow=2.*
```

A.3 REPRODUCIBILITY CONFIGURATION

- a. **Train/Test Split:** 80/20 (stratified by label).
- b. **Cross-Validation:** 10-fold on training set.
- c. **Random Seed:** Fixed across all runs for consistency.
- d. **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score, ROC-AUC.
- e. **Hardware Utilization:** CUDA acceleration (PyTorch backend).

A.4 EXECUTION STEPS

- a. Install dependencies using (`conda env create -f environment.yml`).
- b. Activate environment via (`conda activate thesis-aigt`).
- c. Open JupyterLab and configure dataset paths in the first cell of each notebook.
- d. Execute cells sequentially to reproduce training, validation, and testing.
- e. Results (ROC curves, metrics, and confusion matrices) are auto-saved to (`/outputs/`).

A.5 CODE PACKAGE

A digital archive (`Thesis_Code_Package.zip`) accompanies this thesis and includes:

- a. Scenario 1 and Scenario 2 notebooks (`.ipynb`)
- b. (`environment.yml`) file
- c. Trained model artifacts (`.pkl`)
- d. (`README.md`) for quick replication

All scripts are released under an **MIT License** for academic reuse.

APPENDIX B

DATASETS AND SPLITS

B.1 DATASET OVERVIEW

This research utilized three balanced datasets designed for differentiating between human-written and AI-generated academic or general text, as shown in the Table B.1.

Table B.1: Overview of Dataset Composition and Class Balance.

Dataset	Domain	AI Samples	Human Samples	Total	Balance
AI-GA	Academic abstracts	14,331	14,331	28,662	Balanced
HWAI	Wikipedia articles	22,081	22,081	44,162	Balanced
HAGT-1M	Mixed scholarly & general texts	250,000	250,000	500,000	Balanced

All datasets are English-language, anonymized, and used strictly under research/ethical guidelines.

B.2 PREPROCESSING PIPELINE

1. Text cleaning (HTML/special-character removal, whitespace normalization).
2. Lowercasing and token normalization.
3. Stop-word filtering; removal of extremely frequent domain-generic tokens when necessary (e.g., “paper”, “study” for AI-GA).
4. Vectorization for Scenario 2 (TF-IDF, 1–2 grams, max_features = 10,000) and tokenizer-specific preparation for Scenario 1 (SBERT/RoBERTa).
5. Integrity checks to preserve 1:1 class balance after preprocessing.

B.3 SPLIT MECHANISM

- **Train/Test:** 80/20 stratified split for each dataset.
- **Model Selection:** 10-fold cross-validation on the training portion.
- **Metrics:** Accuracy, Precision, Recall, F1-Score, ROC-AUC (reported per dataset).

B.4 DATA ETHICS AND LICENSING

All corpora are publicly available or openly documented, contain no personally identifiable information, and were handled in compliance with institutional ethics and applicable licenses.

B.5 REPRESENTATIVE CROSS-SECTIONS

The following figures present sample visualizations from each dataset, demonstrating typical contrasts between AI-generated and human-written texts used for training and evaluation. Only short excerpts are displayed to ensure brevity and compliance with data ethics. Representative examples of each dataset are provided in Figures B.1–B.3.

```
df_ai = df[df["label"] == 1]
df_human = df[df["label"] == 0]
print("----- : AI Text Sample : -----")
print(df_ai.iloc[0]["abstract"])
print("\n\n")
print("----- : Human Text Sample : -----")
print(df_human.iloc[0]["abstract"])
```

----- : AI Text Sample : -----

"Epidemiological and Clinical Characteristics of Childhood Pandemic 2009 H1N1 Virus Infection: An Observational Cohort Study" examines the clinical features, disease manifestations, and outcomes in children who were infected with pandemic-associated influenza A (H1N1) virus during 2009. During this season it was estimated that between 11%-21% of symptomatic global population suffered from this seasonal outbreak. Despite widespread availability of both prophylactic as well as treatment measures like neuraminidase inhibitors available for disease control; major morbidity occurred amongst pediatric population due to different epidemiological profiles compared to adults associated with the novel pandemic strain. The objective of this prospective observational cohort study is to investigate epidemiologically related factors along with clinical characteristics among hospitalized pediatric patients towards ascertaining the severity markers associated with severe cases leading to worse prognosis in an attempt towards identifying targets for improved management protocols. The findings from such investigations will reflect upon greater insight on childhood infection induced by H1N1 virus paving a path for future preventive therapeutics against significant worldwide events involving novel emerging viruses.

----- : Human Text Sample : -----

OBJECTIVE: This retrospective chart review describes the epidemiology and clinical features of 48 patients with culture-proven *Mycoplasma pneumoniae* infections at King Abdulaziz University Hospital, Jeddah, Saudi Arabia. METHODS: Patients with positive *M. pneumoniae* cultures from respiratory specimens from January 1997 through December 1998 were identified through the Microbiology records. Charts of patients were reviewed. RESULTS: 48 patients were identified, 33 (82.5%) of whom required admission. Most infections (92.5%) were community-acquired. The infection affected all age groups but was most common in infants (32.5%) and pre-school children (22.5%). It occurred year-round but was most common in the fall (35%) and spring (30%). More than three-quarters of patients (77.5%) had comorbidities. Twenty-four isolates (60%) were associated with pneumonia, 14 (35%) with upper respiratory tract infections, and 2 (5%) with bronchitis. Cough (82.5%), fever (75%), and malaise (58.8%) were the most common symptoms, and crepitations (60%), and wheezes (48%) were the most common signs. Most patients with pneumonia had crepitations (79.2%) but only 25% had bronchial breathing. Immunocompromised patients were more likely than non-immunocompromised patients to present with pneumonia (8/9 versus 16/31, $P = 0.05$). Of the 24 patients with pneumonia, 14 (58.3%) had uneventful recovery, 4 (16.7%) recovered following some complications, 3 (12.5%) died because of *M. pneumoniae* infection, and 3 (12.5%) died due to underlying comorbidities. The 3 patients who died of *M. pneumoniae* pneumonia had other comorbidities. CONCLUSION: our results were similar to published data except for the finding that infections were more common in infants and preschool children and that the mortality rate of pneumonia in patients with comorbidities was high.

Figure B.1: AI-GA Dataset Cross-Section — Sample AI vs. Human Abstracts.

```
df_ai = df[df["Class"] == 1]
df_human = df[df["Class"] == 0]
print("----- : AI Text Sample : -----")
print(df_ai.iloc[0]["Text"])
print("\n\n")
print("----- : Human Text Sample : -----")
print(df_human.iloc[0]["Text"])
```

----- : AI Text Sample : -----

The Boat Race 2000 was the annual rowing competition between the universities of Oxford and Cambridge in England. Held on April 2nd, 2000, the race took place on the River Thames in London. Oxford emerged victorious, beating Cambridge by seven lengths. The Oxford team, led by President Matthew Dundas, showed remarkable speed and dominance throughout the race. The victory marked Oxford's fourth consecutive win in the competition, solidifying their status as a powerhouse in collegiate rowing.

----- : Human Text Sample : -----

Mississippi Highway 403 and MS 15 in Mathiston, and it travels north away from the town. It then turns east near Natchez Trace Parkway and then travels below it. MS 403 ends at the intersection of Clarkson Road and Old US 82 near a county highway maintenance barn. The route was designated by 1960, after a paved road extending from Mathiston was constructed around 1956. An extension northwards to MS 50 and MS 341 existed from 1960 to 1967.

Figure B.2: HWA1 Dataset Cross-Section — Sample AI vs. Human Articles.

```

df_ai = df[df['label'] == 1]
df_human = df[df['label'] == 0]
print("----- : AI Text Sample : -----")
print(df_ai.iloc[0]['text'])
print("\n\n")
print("----- : Human Text Sample : -----")
print(df_human.iloc[0]['text'])

```

----- : AI Text Sample : -----

Ultimately, the decision as to whether students should work in schools or their homes depends on many factors and must be made on an individual basis.

When students work in a school setting, they benefit from frequent social interactions and the varied perspectives of their classmates. This can be particularly beneficial for students who are more independent and self-motivated and need less guidance. Social interactions, academic improvement, and peer support are the main choices that need to be weighed in this decision. They can utilize their differences to collaborate closely on projects, which can in turn lead to better academic performance. Students are able to work in a comfortable and distraction-free environment.

When considering whether students should work in schools or their homes, many factors must be taken into consideration. At the same time, students can maintain contact with their peers and teachers online through video conferencing to ensure they do not miss out on the social connections while also avoiding potential problems of in-person interaction.

However, working from home has its own set of benefits. Both options offer advantages, and it is important to assess a student's particular needs and their preferences before making a decision. At the same time, they can build relationships with peers and have support from each other in a positive and safe environment.

----- : Human Text Sample : -----

Even though lots of students like to work at home, I strongly believe its better to work at schools.

Students should have friends in there lives to be more social in the world. One day I saw a student frustrated with work so I decided to help him in America Civil War. In my science class we have projects to do we get in a group of friends to get a good grade and we see a student work by him self. This shows that in schools we should be with other students to bring in ideas, ways that others can't see.

In school you can be well taught and understand better when people like teachers, or other students help them. Those were the things we both like too he was my first friend in that school. I am writing about should students work in schools or their homes. Then one of the kids came up to me and ask what do you like I told him what I love like animals, sports. For example, In history there are students that don't understand they would just sit in their seats and not ask for help. When we got our projects back we got an A and the student got a B. Its best for students to be in school rather than be home school. Students should be in schools because they should be comfortable when meeting, talking with other students because they might have things in common.

At school students can help each other out in class if they dont understand something. When we took our test we both got A's. Most people think its best for kids to be in school than in their homes because it can help them in their social life, make friends in there world instead of being lonely. When I was younger I use to switch schools to schools then I went to a school called Generic_School I was the only new kid in school I had no friends I was the quite in class, then one day I join a club called book club we just read books and talked about what happen in them. In my class students can be helped by fellow classmate or the teacher.

During school projects students can be in partners or in groups to get a good grade

Figure B.3: HAGT-1M Dataset Cross-Section — Sample AI vs. Human Articles.

APPENDIX C

PYTHON CODE IMPLEMENTATION

The main Python implementations created for the research study called "Ensuring Academic Integrity: Transformer-based and Ensemble Machine Learning Approaches for Differentiating Human-written and AI-generated Text" are described in this appendix.

The frameworks were applied together and were experimentally validated on three balanced datasets (AI-GA, HWAI, and HAGT-1M) during the same experiments.

The experiments were conducted entirely on Python 3.10 along with the libraries of PyTorch, scikit-learn, transformers, xgboost, and Matplotlib.

The code snippets below demonstrate the typical implementations of each framework; the same structures were subsequently used for the other datasets with the sole change being the input file path.

C.1 SCENARIO 1 – TRANSFORMER-BASED DETECTION FRAMEWORK

The framework at hand uses Sentence - BERT (SBERT) and RoBERTa embeddings in the process of capturing deep contextual semantics and syntactic dependencies.

Two classifiers were utilized for performance evaluation, namely, Logistic Regression (LR) and a Feedforward Neural Network (FNN). The approach consists of four major steps:

- a. **Text Embedding:** The creation of high-dimensional contextual vectors through SBERT and RoBERTa.
- b. **Model Training:** Training of Logistic Regression and FNN models on the embedding vectors.
- c. **Evaluation:** Computation of Accuracy, Precision, Recall, F1 Score, and ROC-AUC metrics.
- d. **Visualization:** The performance of the classification is shown through the plots of the Confusion Matrix and the ROC Curve.

```

# ===== SCENARIO 1: TRANSFORMER-BASED DETECTION (AI-GA EXAMPLE) =====
# Sentence-BERT and RoBERTa embeddings with Logistic Regression and FNN.

import pandas as pd, numpy as np, torch, torch.nn as nn, torch.optim as optim
import matplotlib.pyplot as plt, seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score, KFold
from sentence_transformers import SentenceTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import (
    accuracy_score, precision_recall_fscore_support, roc_auc_score,
    classification_report, confusion_matrix, ConfusionMatrixDisplay, roc_curve
)

# --- Load and preprocess dataset ---
DATA_PATH = "AI-GA dataset.csv"
df = pd.read_csv(DATA_PATH).drop_duplicates(["title", "abstract"]).dropna()
train_texts, test_texts, y_train, y_test = train_test_split(
    df["abstract"], df["label"], test_size=0.2, random_state=42,
    stratify=df["label"]
)

# --- Define Feedforward Neural Network ---
class FeedforwardNN(nn.Module):
    def __init__(self, input_dim):
        super().__init__()
        self.fc1, self.fc2, self.fc3 = nn.Linear(input_dim, 512),
            nn.Linear(512, 128), nn.Linear(128, 2)
        self.relu = nn.ReLU()
    def forward(self, x):
        x=self.relu(self.fc1(x)); x=self.relu(self.fc2(x)); return self.fc3(x)

```

```

# --- FNN Training and Evaluation Functions ---
def train_fnn(X_train, y_train, X_val, y_val, dim, epochs=10, batch=32):
    model, opt = FeedforwardNN(dim), optim.Adam(FeedforwardNN(dim).parameters()),
                                                lr=1e-3)

    loss_fn = nn.CrossEntropyLoss()
    for epoch in range(epochs):
        model.train()
        perm = torch.randperm(X_train.size(0))
        for i in range(0, len(perm), batch):
            idx = perm[i:i+batch]; xb, yb = X_train[idx], y_train[idx]
            opt.zero_grad(); out = model(xb); loss = loss_fn(out, yb);
            loss.backward(); opt.step()
        model.eval()
        with torch.no_grad():
            val_out = model(X_val)
            val_acc = accuracy_score(y_val, val_out.argmax(1))
        print(f"Epoch {epoch+1}/{epochs} - Validation Accuracy: {val_acc:.3f}")
    return model

def evaluate_lr(X_test, y_test, clf):
    pred = clf.predict(X_test); prob = clf.predict_proba(X_test)[:,:1]
    acc = accuracy_score(y_test, pred)
    pre, rec, f1, _ = precision_recall_fscore_support(y_test, pred, average="binary")
    auc = roc_auc_score(y_test, prob)
    print(f"ACC={acc:.3f} PRE={pre:.3f} REC={rec:.3f} F1={f1:.3f}
    AUC={auc:.3f}")

```

```

# --- Transformer Embedding and Classification ---
models = {
    "SBERT": "distilbert-base-nli-stsb-mean-tokens",
    "RoBERTa": "sentence-transformers/roberta-base-nli-stsb-mean-tokens"
}

for name, model_id in models.items():
    print(f"\n===== {name} MODEL =====")
    encoder = SentenceTransformer(model_id)
    X_train = encoder.encode(train_texts.tolist(), show_progress_bar=True)
    X_test = encoder.encode(test_texts.tolist(), show_progress_bar=True)

    # Logistic Regression
    lr = LogisticRegression(max_iter=1000, random_state=42)
    lr.fit(X_train, y_train)
    cv = cross_val_score(lr, X_train, y_train, cv=5, scoring="accuracy")
    print(f"{name} LR Mean CV Accuracy: {cv.mean():.4f}")
    evaluate_lr(X_test, y_test, lr)

    # Feedforward Neural Network
    Xt, Xv = torch.tensor(X_train, dtype=torch.float32),
                  torch.tensor(X_test, dtype=torch.float32)
    yt, yv = torch.tensor(y_train.values, dtype=torch.long),
                  torch.tensor(y_test.values, dtype=torch.long)
    fnn = train_fnn(Xt, yt, Xv, yv, X_train.shape[1])

```

For the **HWAI** and **HAGT-1M** datasets, the identical structure is recycled just by changing the (`DATA_PATH`) variable. The training and testing phases are done in the same way, thus providing methodological consistency throughout the different scenarios.

C.2 SCENARIO 2 – ENSEMBLE LEARNING FRAMEWORK

The second framework is a soft-voting ensemble of multiple classical machine-learning algorithms that improves interpretability and robustness. Text is represented using TF-IDF vectorization, and seven classifiers (LR, SVM, XGBoost, Random Forest, Extremely Randomized Trees, AdaBoost, and SGD) are combined to yield the final prediction. Implementation Highlights:

1. **Feature Extraction:** TF-IDF (10 000 features, bigrams).

2. **Base Classifiers:** Logistic Regression, SVM, XGBoost, Random Forest, Extremely Randomized Trees, AdaBoost, SGD.
3. **Soft Voting Ensemble:** Combining probability estimates for balanced decision fusion.
4. **Evaluation:** Accuracy, Precision, Recall, F1, ROC-AUC, Confusion Matrix, and 10-Fold Cross-Validation.

```
# ===== SCENARIO 2: ENSEMBLE LEARNING (AI-GA EXAMPLE) =====

import pandas as pd, numpy as np, seaborn as sns, matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, StratifiedKFold,
cross_val_score
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression, SGDClassifier
from sklearn.svm import SVC
from sklearn.ensemble import (
    RandomForestClassifier, ExtraTreesClassifier, AdaBoostClassifier,
    VotingClassifier
)
from xgboost import XGBClassifier
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    roc_auc_score, roc_curve, auc, confusion_matrix, ConfusionMatrixDisplay
)

# --- Data Loading and Preprocessing ---
df = pd.read_csv("AI-GA dataset.csv").drop_duplicates(["title", "abstract"]).dropna()
X_text, y = df["abstract"], df["label"]

# --- TF-IDF Vectorization ---
tfidf = TfidfVectorizer(max_features=10000, ngram_range=(1,2),
                        stop_words="english")
X = tfidf.fit_transform(X_text)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
                                                    random_state=42, stratify=y)
```

```

# --- Define Models ---
lr = LogisticRegression(max_iter=2000, solver="liblinear")
svm = SVC(probability=True, random_state=42)
xgb = XGBClassifier(eval_metric="logloss", random_state=42)
rf = RandomForestClassifier(n_estimators=500, random_state=42)
ert = ExtraTreesClassifier(n_estimators=300, bootstrap=True, random_state=6)
ada = AdaBoostClassifier(n_estimators=200, learning_rate=0.5, random_state=1)
sgd = SGDClassifier(max_iter=5000, loss="modified_huber", random_state=42)

# --- Soft Voting Ensemble ---
ensemble = VotingClassifier(
    estimators=[("lr", lr), ("svm", svm), ("xgb", xgb), ("rf", rf),
                ("ert", ert), ("ada", ada), ("sgd", sgd)],
    voting="soft", n_jobs=-1
)
ensemble.fit(X_train, y_train)

```

```

# --- Evaluation ---
def evaluate(model, X_test, y_test, name):
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:,1]
    acc = accuracy_score(y_test, y_pred)
    pre = precision_score(y_test, y_pred)
    rec = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    auc_ = roc_auc_score(y_test, y_prob)
    print(f"[{name}] ACC={acc:.3f} PRE={pre:.3f} REC={rec:.3f} F1={f1:.3f} AUC={auc_:.3f}")
    ConfusionMatrixDisplay.from_predictions(y_test, y_pred, cmap="Blues")
    plt.title(f"Confusion Matrix - {name}")
    plt.show()

for model_name, model in [("Logistic Regression", lr), ("SVM", svm),
                          ("XGBoost", xgb), ("Random Forest", rf),
                          ("Extra Trees", ert), ("AdaBoost", ada), ("SGD", sgd),
                          ("Voting Ensemble", ensemble)]:
    model.fit(X_train, y_train)
    evaluate(model, X_test, y_test, model_name)

# --- 10-Fold Cross Validation for Ensemble ---
cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
cv_scores = cross_val_score(ensemble, X, y, cv=cv, scoring="accuracy")
print(f"10-Fold CV Accuracy: {cv_scores.mean()*100:.2f} ± {cv_scores.std()*100:.2f}%")

```

On the **HWAI** and **HAGT-1M** datasets, the identical script was performed unaltered, with the only exception being the file path of the dataset. The performance measured on the different datasets validated the proposed ensemble framework's consistency and generalization capabilities.